
Evaluating the Generalizability of Deepfake Detection Models: A Comparative Analysis of GAN and Diffusion-Based Generated Content

By

Md. Tarek Hasan (0122410019)

Submitted in partial fulfilment of the requirements
of the degree of Master of Science in Computer Science and Engineering

August 2, 2025



DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
UNITED INTERNATIONAL UNIVERSITY

Abstract

The rapid advancement of deepfake technology has significantly increased the realism and accessibility of synthetic media. Emerging techniques such as diffusion-based models and Neural Radiance Fields (NeRF), along with improvements in traditional Generative Adversarial Networks (GANs), have enabled the sophisticated generation of deepfake videos, posing growing threats to biometric security and trust. In parallel, detection methods have advanced through innovations in Transformer-based architectures, contrastive learning, and other deep learning approaches. Yet, this progress continues to play out within a persistent cat-and-mouse dynamic between generation and detection. In this work, we present a comprehensive empirical evaluation of state-of-the-art deepfake detection methods, alongside a human subject study focused on identifying deepfakes, using curated stimuli set generated by cutting-edge deepfake synthesis techniques. Unlike prior efforts, our study establishes a benchmark that captures the challenges posed by the latest generation methods in realistic settings. Our findings expose a critical vulnerability: both leading detection models and human evaluators struggle when confronted with high-quality, modern deepfakes. To address this gap, we introduce a multimodal detection framework that incorporates both audio and visual modalities, enhancing the robustness of detection systems in cross-modal scenarios. Our methodology includes evaluating performance across diverse conditions—such as different resolutions and clip lengths—and comparing unimodal versus multimodal fusion strategies. Extensive experimentation highlights the urgent need to refine detection models to keep pace with rapidly evolving generative techniques. By establishing a rigorous benchmark and revealing current limitations, our study offers a timely foundation for developing more robust and future-ready deepfake detection systems. Our results demonstrate that incorporating the audio modality alongside video consistently improves detection performance, underscoring the value of multimodal analysis for robust generalization. Notably, the proposed multimodal framework—evaluated on the FakeAVCeleb and AV-Deepfake1M datasets—achieved superior performance across all tested conditions, with early fusion yielding the highest AUC and precision, and cross-modal attention demonstrating particular effectiveness under low-resolution scenarios.

Acknowledgements

This journey began in January 2024 when I took my first steps into the Master of Science in Computer Science and Engineering program. After completing the admission procedures, I reached out to Prof. Swakkhar Shatabda, who had been my undergraduate thesis supervisor and a guiding force in my academic life. During my undergraduate years, I worked on Natural Language Processing under his supervision. However, I felt the urge to explore a new dimension in my research journey for my master's degree and expressed my interest in diving into Computer Vision.

Before embarking on this academic path, I would like to extend my heartfelt gratitude to Prof. Dr. Md. Motaharul Islam for his encouragement and unwavering support, which motivated me to pursue the master's degree. His guidance and belief in my potential gave me the confidence to take this important step.

Prof. Swakkhar, ever encouraging and insightful, introduced me to two research papers, suggesting that I read and reflect on them. After immersing myself in the contents, one topic resonated with me, and that marked the beginning of my deepfake research. To guide me further, Prof. Swakkhar connected me with Dr. Sanjay Saha, the first author of the selected paper and an expert in deepfake research. Little did I know then that this connection would turn into an invaluable mentorship. Dr. Sanjay, who had delved into deepfake research during his Ph.D., became a guiding light on my path from deepfake generation to detection. His advice and mentorship were nothing short of life-saving during the challenging phases of my work.

While navigating the world of deepfakes, I was fortunate enough to secure a research internship at the Multimedia Research Lab of Prof. Dr. Terence Sim at the National University of Singapore. Working under Prof. Dr. Terence Sim, a pioneer in deepfake research, Dr. Sanka Rasnayaka, and Dr. Shaojing Fan, I delved deeper into not only deepfake detection but also biometric authentication. The exposure and experience were truly enriching, and I owe a debt of gratitude to Prof. Dr. Terence Sim for his invaluable support, both academically and personally, during my stay in Singapore. I am especially grateful for the financial and computational support he provided for this research, which made my work more feasible and impactful. Adjusting to life there wasn't easy, but Dr. Sanjay's constant encouragement and mentorship made it manageable.

Throughout this journey, Prof. Swakkhar continued to inspire and push me to meet deadlines, reinforcing the importance of dedication and perseverance. I remain deeply grateful to him for introducing me to Dr. Sanjay, who turned out to be one of the best mentors I could ever ask for. The unwavering support of both mentors has shaped not just this thesis but my entire academic outlook. I would also like to extend my sincere gratitude to Assistant Professor Nahid Hossain for his constant administrative support and for being a source of motivation during challenging times.

None of this would have been possible without the endless support and patience of my beloved wife. She gave me the space and encouragement to immerse myself in my research from dawn to dusk, never once complaining about the time I dedicated to my work instead of to our life together. She truly is the greatest blessing in my life, and I am endlessly thankful to her. I am equally grateful to my parents and sister, who have always been my pillars of strength. The time that should have been theirs, I often dedicated to this thesis, and they never once complained, allowing me to pursue my passion without any added stress.

Finally, I owe everything to the Almighty Allah for making this journey possible. Through all the ups and downs, His guidance and blessings kept me moving forward. Alhamdulillah.

Publication List

The primary contributions of this research have been submitted for peer review and are currently under consideration in a reputed conference, while a related journal article—containing partial contributions from this thesis—has already been published, as detailed below:

Conference Paper

1. *Deepfake Synthesis vs. Detection: An Uneven Contest* – currently under review.

Journal Article

1. *AniFaceDiff: Animating Stylized Avatars via Parametric Conditioned Diffusion Models* – published in Pattern Recognition (Elsevier, Q1, Impact Factor: 7.6). This article includes partial contributions from the present thesis.

Table of Contents

Table of Contents	v
List of Figures	vi
List of Tables	x
1 Introduction	1
1.1 Research Overview and Context	1
1.2 Motivation and Research Significance	2
1.3 Research Objectives and Goals	3
1.4 Methodological Approach	3
1.5 Key Contributions and Findings	4
1.6 Structure and Organization of the Thesis	5
2 Background and Literature Review	6
2.1 Preliminaries	6
2.1.1 Understanding Deepfakes	6
2.1.2 Classification of Deepfake Techniques	7
2.1.3 Models and Techniques for Deepfake Generation	7
2.1.4 State-of-the-Art Deepfake Detection Architectures	9
2.1.5 Challenges in Multimodal Deepfake Detection	10
2.2 Related Works	10
2.2.1 Summary of Gaps and Research Questions	11
2.3 Gap Analysis	14
2.4 Summary	15
3 Methodology and Experimental Design	16
3.1 Requirement Analysis	16
3.2 Experimental Design	17
3.2.1 Data Preparation: Deepfake Synthesis	18
3.2.2 Deepfake Detection Algorithm Consolidation	21
3.2.3 Human Level Detection	25
3.2.4 Multimodal Integration	30

3.3	Summary	37
4	Implementation, Evaluation, and Results	38
4.1	Experimental Environment and Setup	38
4.2	Hyperparameter Settings	38
4.3	Performance Evaluation	39
4.3.1	Synthetic Video Generation and Dataset Preparation	39
4.3.2	Unimodal Deepfake Detection Analysis	39
4.3.3	Benchmarking and Performance Profiling of Detection Models	46
4.3.4	Impact of Video Conditions on Detection Performance	49
4.3.5	Multimodal Fusion Strategies and Detection Results on Novel Sce- narios	50
4.4	Summary	51
5	Conclusion and Future Directions	53
5.1	Summary	53
5.2	Limitation	53
5.3	Future Work	54
	References	61

List of Figures

1.1	A chronological timeline showcasing the evolution of talking head generation models (top, yellow) and deepfake detection methods (bottom, green). The diagram highlights key generative models such as FaceVid [1], TPSMM [2], VASA-1 [3], and SyncTalk [4], alongside notable detection techniques like MesoNet [5], Xception [6], SRM [7], and NPR [8]. While some detection models predate 2021, this timeline emphasizes major developments from 2021 to 2024, illustrating the ongoing arms race between deepfake generation and detection technologies.	2
1.2	Overview of the research methodology pipeline for deepfake detection and analysis. The process begins with dataset generation, followed by parallel evaluation of existing detection models and a human visual deepfake detection study. Insights from both analyses inform benchmarking and qualitative assessment, leading to the development of a multimodal detection framework, which is subsequently validated.	4
2.1	Illustration of Deepfake Manipulation Techniques: Face Swapping and Expression Reenactment. The left column outlines the conceptual manipulation workflow, while the right column presents generated media samples. In face swapping, Person A’s facial identity is transferred onto Person B’s body, maintaining the latter’s expressions [9]. In expression reenactment, Person A’s facial identity is retained, but Person B’s expressions are transferred, allowing Person A to mimic Person B’s emotional or speech-related movements [10].	8

- 2.2 Conceptual Comparison of Generative Models Used in Deepfake Synthesis. This diagram illustrates the high-level workflows of three foundational generative models: Generative Adversarial Networks (GANs), Diffusion Models, and Neural Radiance Fields (NeRFs). All models are shown generating facial outputs to highlight their application in deepfake media synthesis. GANs utilize an adversarial training process between a generator and a discriminator. Diffusion models progressively generate images by reversing a noise-adding process. NeRFs represent 3D scenes by learning view-dependent radiance, enabling photorealistic novel-view synthesis from a learned 3D face representation. 9
- 3.1 Sequence of frames from videos generated using the deepfake synthesis methods discussed in this paper. The figure showcases high-quality and realistic deepfake videos, demonstrating realistic facial movements that are challenging to detect with casual inspection. Methods depicted in the figure are respectively: AniFaceDiff [11], FaceFusion [12], FaceVid [1], FADM [13], HyperReenact [14], SyncTalk [4], TPSMM [2], and VASA-1 [3]. 21
- 3.2 Demographic breakdown of participants in the deepfake detection survey. The majority of participants reported high prior AI experience (73.4%) and held at least a bachelor's degree, with over half possessing a master's degree. The largest age group was 25–34 years, indicating a relatively young and technologically proficient participant pool. 27
- 3.3 Overview of the multimodal deepfake detection framework. The pipeline begins with feature extraction, where video frames are sampled and audio is extracted from the input video. Temporal modeling is performed using TimeSformer for video and Wav2Vec 2.0 for audio to obtain modality-specific embeddings. Three fusion strategies are explored: (i) Early Fusion, where audio features are projected to match video dimensionality and concatenated before classification; (ii) Cross-Modal Attention Fusion, a form of early fusion enhanced with multi-head attention, where interactions between audio and visual features are learned and integrated into a fused representation before classification; and (iii) Late Fusion, where separate modality-specific classifiers are used and their outputs are aggregated. 30

- 3.4 Illustration of the proposed Cross-Modal Attention mechanism for multi-modal deepfake detection. Visual features extracted from TimeSformer act as Queries, while audio features from Wav2Vec 2.0 serve as both Keys and Values. This setup allows the video modality to dynamically attend to the most informative segments of the audio stream. As a result, audio cues guide the refinement of visual representations, enabling the model to highlight inconsistencies—such as lip-sync mismatches or emotion desynchronization—that are commonly found in deepfakes. 36
- 4.1 Cross-model detection performance (AUC scores) of the four best-performing automated deepfake detection algorithms compared to human participants, evaluated across generative models. Humans consistently outperform all automated methods across all generators, particularly excelling against An-iFaceDiff [11] and HyperReenact [14]. Humans demonstrate a strong capability to detect deepfakes generated by state-of-the-art (SOTA) models such as SyncTalk [4], DreamTalk [15], Diffused Heads [16], and VASA-1 [3], highlighting their perceptual advantage over current automated detection systems. The results highlight the current limitations of detection algorithms in generalizing across diverse deepfake generation techniques.
40
- 4.2 Δ AUC (High-Res – Low-Res) for each detection model, representing robustness to video resolution changes. Each AUC value is the average performance across multiple generative models (e.g., different deepfake sources). Positive Δ values indicate better performance on high-resolution videos. Xception [6] and Human detectors show the largest gains (>8 AUC points), while EfficientNet-B4 [17] and RECCE [18] perform worse on high-res inputs. The overall average Δ AUC across all models is $+1.90$, suggesting that high-resolution videos modestly improve detection performance on average.
41
- 4.3 Aggregate impact of prior AI experience on fake video classification accuracy. This figure shows the average classification error across participant groups with different levels of AI experience. Participants with low AI experience had the highest overall error rates, while those with high experience performed significantly better. These results indicate a clear negative correlation between AI expertise and misclassification of fake videos at the group level. 47

4.4 Individual classification responses to fake videos by AI experience level.
This figure captures how individual participants rated individual fake videos,
broken down by their level of prior AI experience. Participants with low
AI experience were more likely to misclassify fake content as real, whereas
those with high experience more frequently labeled videos as ‘Definitely
computer generated.’ The results suggest that AI expertise improves de-
tection accuracy on a video-by-video basis. 47

List of Tables

- 2.1 Comparative Summary of Existing Review Articles on Deepfake Generation vs Detection: Focus Areas, Methodologies, and Research Contributions
13
- 2.2 Addressing Key Gaps in Existing Deepfake Literature: Methodological Enhancements and Focus Areas
14
- 3.1 Comprehensive summary of various deepfake generation techniques, detailing the method, datasets and number of images/videos used, and key highlights for each approach.
22
- 3.2 Summary of different deepfake detection methods, including their publication details, datasets, techniques, model types, and key highlights.
26
- 4.1 Comparison of Detection and Generation Performance (AUC and AP values) on 5-Second Videos for Various Deepfake Detection Models (ROWS) and Generators (COLUMNS). The bolded values highlight the best overall detection model across rows and the most challenging deepfake generator across columns.
42
- 4.2 Comparison of Detection and Generation Performance (P, Precision and R, Recall Values) on 5-Second Videos for Various Deepfake Detection Models (ROWS) and Generators (COLUMNS). The bolded values highlight the best overall detection model across rows and the most challenging deepfake generator across columns.
42

- 4.3 Comparison of Detection and Generation Performance (d' , Sensitivity Index and C, Decision Criterion Values) on 5-Second Videos for Various Deepfake Detection Models (ROWS) and Generators (COLUMNS). The bolded values highlight the best overall detection model across rows and the most challenging deepfake generator across columns.
43
- 4.4 Comparison of Detection and Generation Performance (AUC and AP values) on 5-second High Resolution Videos for Various Deepfake Detection Models (ROWS) and Generators (COLUMNS). The bolded values highlight the best overall detection model across rows and the most challenging deepfake generator across columns.
43
- 4.5 Comparison of Detection and Generation Performance (AUC and AP values) on 5-second Low Resolution Videos for Various Deepfake Detection Models (ROWS) and Generators (COLUMNS). The bolded values highlight the best overall detection model across rows and the most challenging deepfake generator across columns.
44
- 4.6 Comparison of Detection and Generation Performance (AUC and AP values) on Full Length Videos for Various Deepfake Detection Models (ROWS) and Generators (COLUMNS). The bolded values highlight the best overall detection model across rows and the most challenging deepfake generator across columns.
44
- 4.7 Comparison of Detection and Generation Performance (P, Precision and R, Recall Values) on Full Length Videos for Various Deepfake Detection Models (ROWS) and Generators (COLUMNS). The bolded values highlight the best overall detection model across rows and the most challenging deepfake generator across columns.
45
- 4.8 Comparison of Detection and Generation Performance (d' , Sensitivity Index and C, Decision Criterion Values) on Full Length Videos for Various Deepfake Detection Models (ROWS) and Generators (COLUMNS). The bolded values highlight the best overall detection model across rows and the most challenging deepfake generator across columns.
45

4.9	Average AUC (%) of Deepfake Detection Models Across Varying Video Durations and Resolutions. Each value represents the model’s mean performance across multiple generative methods. Detection performance is evaluated on full-length videos, 5-second clips, and 5-second clips at low and high resolutions. The bolded values highlight the best detection model on each video condition. <i>Note: Human evaluation was not conducted on full-length videos; corresponding value is not available.</i>	49
4.10	Performance comparison of unimodal and multimodal deepfake detection across varying durations, resolutions, and fusion strategies. The table reports AUC, AP, precision, recall, Sensitivity Index (d-prime), and Decision Criterion (C) for both full-length and 5-second video clips. Unimodal (video-only or audio-only) and multimodal (audio-video) configurations are evaluated under high/low resolution conditions. Multimodal fusion strategies include Early Fusion, Late Fusion, and Cross-Modal Attention. While all models show performance degradation on shorter clips, Cross-Modal Attention consistently outperforms other fusion approaches under low-resolution settings, highlighting its effectiveness in learning cross-modal representations when visual quality is limited.	51

Chapter 1

Introduction

This chapter introduces the scope, motivation, methodology, and expected outcomes of the thesis. It outlines the current landscape of deepfake technology, the challenges in detection, and the rationale behind this research.

1.1 Research Overview and Context

The proliferation of deepfake technology has fundamentally transformed the landscape of digital media. With advanced machine learning techniques, highly realistic synthetic content—ranging from images and videos to audio—is being created to manipulate identities or fabricate behaviors. Deepfakes can take several forms, including face swapping, expression or mouth movement reenactment, and full-face synthesis, each capable of deceiving viewers with striking realism [19, 20].

Originally fueled by Generative Adversarial Networks (GANs) [21], deepfake generation has rapidly progressed with the inclusion of Variational Autoencoders (VAEs), diffusion-based models [22, 23], and neural rendering approaches such as Neural Radiance Fields (NeRF) [24]. These modern techniques significantly elevate the fidelity and believability of deepfake content, making them less visually detectable than earlier GAN-based artifacts [25].

While these advancements unlock creative potential in entertainment and media, they simultaneously present serious ethical and societal risks. Deepfakes are increasingly weaponized in disinformation campaigns, identity fraud, and social engineering attacks [26]. Sophisticated talking-head synthesis and face reenactment systems—such as DiffusedHeads [16], DreamTalk [15], AniFaceDiff [11], VASA [3], and SyncTalk [4]—epitomize the growing complexity of synthetic media.

In response, detection models—primarily built upon deep learning architectures such as convolutional neural networks (CNNs), Transformer encoders, and contrastive learning methods—have emerged to identify subtle inconsistencies between real and manipulated content [20]. However, the rapid evolution of generative techniques has outpaced the development of detection systems, resulting in a persistent and escalating arms race between

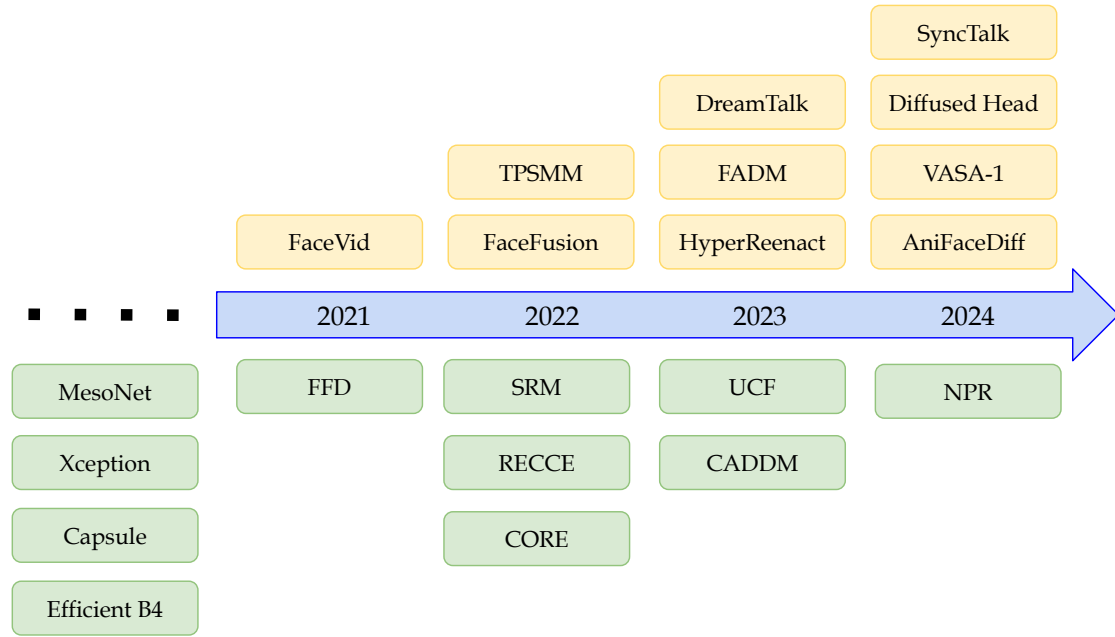


Figure 1.1: A chronological timeline showcasing the evolution of talking head generation models (top, yellow) and deepfake detection methods (bottom, green). The diagram highlights key generative models such as FaceVid [1], TPSMM [2], VASA-1 [3], and SyncTalk [4], alongside notable detection techniques like MesoNet [5], Xception [6], SRM [7], and NPR [8]. While some detection models predate 2021, this timeline emphasizes major developments from 2021 to 2024, illustrating the ongoing arms race between deepfake generation and detection technologies.

generation and detection [27]. This dynamic is illustrated in Figure 1.1, which presents a chronological overview of major developments in both talking head generation and deepfake detection methods from 2021 to 2024.

This thesis identifies this critical imbalance by evaluating the robustness of state-of-the-art detection approaches and investigating the human capacity for deepfake identification—highlighting both technical and perceptual vulnerabilities.

1.2 Motivation and Research Significance

Despite continuous innovation in deepfake detection—from early shallow CNNs like MesoNet [5] to more recent Transformer-based architectures and contrastive learning paradigms [28, 29]—detection models are struggling to keep pace in the generative arms race. Sophisticated synthesis techniques are evolving faster than their countermeasures, enabling deepfakes to bypass even the most advanced detectors.

This growing asymmetry has profound consequences. As manipulated media becomes harder to detect, the authenticity of digital content is increasingly undermined, eroding public trust and enabling the spread of misinformation at scale. Furthermore, as deepfake content begins to mimic real-world scenarios more accurately—such as speech synchro-

nization or emotional mimicry—the potential for abuse intensifies.

Therefore research is driven by the urgent need to assess the gap between deepfake generation and detection, through comprehensive benchmarking and human-centered evaluations that can inform the design of next-generation detection tools.

1.3 Research Objectives and Goals

This thesis aims to address the challenges in deepfake detection through a systematic investigation guided by the following key objectives:

- To empirically assess the effectiveness of state-of-the-art deepfake detection models when evaluated against deepfakes generated using recent and advanced synthesis techniques.
- To benchmark and critically analyze the limitations and vulnerabilities of existing detection approaches through a comprehensive set of experiments across diverse datasets and model architectures.
- To design and conduct a human subject study evaluating individuals’ ability to discern deepfakes, with the goal of uncovering perceptual cues and behavioral patterns that could inform the development of more robust detection systems.
- To investigate whether incorporating the audio modality alongside video improves the performance of deepfake detection systems through empirical evaluation of multimodal fusion strategies.
- To propose informed directions for future research based on insights derived from both machine-based evaluations and human study outcomes.

1.4 Methodological Approach

The methodology involves a comprehensive empirical evaluation framework that tests the performance of several recent detection models on deepfakes generated by cutting-edge synthesis methods. The evaluation spans both machine-learning-based detectors and human participants.

First, we curate and use multiple datasets of deepfake videos generated from diverse pipelines, including GANs, diffusion models, and neural rendering. We then benchmark detection models ranging from CNN-based architectures to Transformer and contrastive learning models. In parallel, a human study is conducted where participants are asked to classify real versus fake videos. Results from both experiments are analyzed to identify weaknesses in current detection strategies and suggest directions for improvement.

An overview of the complete experimental pipeline, including dataset generation, model and human evaluations, benchmarking, the development of a multimodal detection framework, and final validation, is illustrated in Figure 1.2.

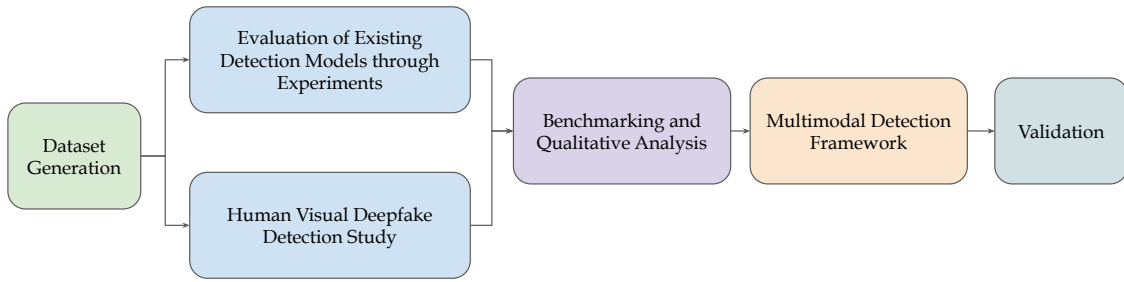


Figure 1.2: Overview of the research methodology pipeline for deepfake detection and analysis. The process begins with dataset generation, followed by parallel evaluation of existing detection models and a human visual deepfake detection study. Insights from both analyses inform benchmarking and qualitative assessment, leading to the development of a multimodal detection framework, which is subsequently validated.

1.5 Key Contributions and Findings

Our findings reveal critical insights into the current limitations of deepfake detection systems and offer practical guidance for future research directions. Rather than proposing a new state-of-the-art detection model, our focus was to systematically investigate whether integrating additional modalities—specifically audio—enhances detection robustness. Through rigorous experimentation with a multimodal detection framework, we demonstrate that incorporating the audio modality alongside video provides consistent performance gains over unimodal baselines. This experimental evidence serves as the central contribution of our proposed approach.

Our key contributions are as follows:

- **Benchmark and Empirical Evaluation:** We establish a rigorous benchmark through a comprehensive empirical analysis of modern deepfake generation and detection techniques, using both unimodal and multimodal inputs. Our results provide valuable insights into detection performance across diverse generative methods and conditions.
- **Human Evaluation and Insight Integration:** A large-scale human detection study was conducted to evaluate perceptual sensitivity to deepfakes. The results not only affirm human superiority in certain detection scenarios but also reveal how prior AI experience impacts accuracy and confidence. These insights inform future human-in-the-loop system design.
- **Experimental Proposal for Multimodal Detection:** We present an experimental validation of a unified multimodal detection framework combining video and audio modalities. The framework incorporates fusion strategies (early, late, and cross-modal attention) and leverages pretrained models (TimeSformer and Wav2Vec2.0). The results consistently show that adding audio improves detection performance—an insight we propose as a guiding direction for future multimodal deepfake research.

This work aims to inform and steer the design of future detection systems by identifying where current methods fall short and providing evidence-backed recommendations, especially in the context of multimodal detection.

1.6 Structure and Organization of the Thesis

This thesis is organized into five chapters. Chapter 1 provides an introduction to the research problem, outlining the motivation, objectives, and methodology employed in the study. It also highlights the key contributions and outcomes of the research. Chapter 2 presents the background and literature review, covering the fundamental concepts related to deepfake generation and detection, including GANs, diffusion models, and multimodal approaches. It also discusses existing detection techniques and identifies the research gaps that this study aims to address. Chapter 3 details the methodology, including the design and implementation of the deepfake detection framework. It describes the dataset preparation, model selection, human evaluation study, and the multimodal fusion strategies employed to improve detection accuracy. Chapter 4 discusses the implementation and presents the results of the proposed multimodal deepfake detection system. It covers the environment setup, benchmark evaluations of unimodal and multimodal methods, comparative analysis with human performance, and insights from adding audio as a new modality alongside video. Finally, Chapter 5 concludes the thesis by summarizing the key findings, discussing the limitations encountered during the research, and proposing future directions, including the exploration of multimodal deepfake detection and human performance analysis in multimodal settings.

Chapter 2

Background and Literature Review

This chapter introduces foundational concepts, recent developments, and existing literature on deepfake synthesis and detection. It includes necessary background knowledge, a literature review of prior work, a gap analysis, and a summary of this chapter’s insights.

2.1 Preliminaries

In this section, we present essential background concepts and terminology required to understand deepfake synthesis and detection, based on prior foundational research and recent advancements in generative AI.

2.1.1 Understanding Deepfakes

Deepfakes refer to synthetic media—such as images, videos, or audio—that are generated or manipulated using artificial intelligence (AI) techniques, particularly deep learning. These outputs are often so realistic that they become difficult to distinguish from authentic content. While the term “deepfake” often carries negative connotations, the underlying technology has a broad range of applications, both beneficial and harmful.

On the positive side, deepfake technology has found use in the entertainment industry for creating realistic visual effects in movies and video games, and in accessibility tools such as voice synthesis for individuals with speech impairments. These applications demonstrate the creative and inclusive potential of generative AI. However, the technology has also raised serious ethical and societal concerns. Deepfakes have been weaponized for spreading misinformation, conducting identity fraud, impersonating individuals without consent, and violating personal privacy. This dual-use nature of the technology necessitates a deeper understanding and development of robust detection mechanisms.

2.1.2 Classification of Deepfake Techniques

Deepfakes can be broadly categorized into several types based on the manipulation technique employed. One of the most common techniques is face swapping, where the face of one individual (the source) is superimposed onto the body of another individual (the target). This process retains the expressions and facial movements of the target, making the swap appear natural and convincing.

Another prominent type is expression reenactment, which involves altering the facial expressions or mouth movements of a person in a video to mimic those of another individual. In this case, the source provides the expressions, which are then transferred onto the target, enabling the creation of manipulated videos that appear to show the target saying or emoting something they never did.

Figure 2.1 illustrates the conceptual manipulation workflow and visual output of both techniques, showcasing how deepfake systems modify source and target components to achieve realistic synthetic media.

A more advanced form of manipulation is full face synthesis, where an entirely new and lifelike face is generated from scratch using deep learning models such as GANs (Generative Adversarial Networks) or diffusion models. These synthetic faces often exhibit a high degree of realism and can be indistinguishable from genuine human faces, posing challenges for both human observers and detection algorithms.

2.1.3 Models and Techniques for Deepfake Generation

The rapid advancement of deep learning has led to the development of powerful generative models capable of synthesizing highly realistic media content. These models form the foundation of modern deepfake generation techniques, enabling the creation of synthetic videos, images, and audio that closely resemble authentic human behavior. Over time, the field has evolved from early generative methods to more sophisticated architectures that offer improved fidelity, temporal consistency, and control. This section elaborates on the principal categories of models used in deepfake generation: Generative Adversarial Networks (GANs), Diffusion Models, and Neural Radiance Fields (NeRFs).

Generative Adversarial Networks (GANs) [21] are among the earliest and most widely used architectures for deepfake synthesis. A GAN consists of two neural networks—a generator and a discriminator—that compete in an adversarial training setup. The generator produces synthetic content, such as facial images or video frames, while the discriminator evaluates whether a given sample is real or fake. Through this adversarial process, the generator gradually learns to create outputs that are indistinguishable from real data. GANs have been extensively applied in face swapping, identity manipulation, and lip-syncing tasks. Their efficiency and visual quality have made them a staple in many early deepfake generation systems. However, they are often sensitive to hyperparameters and may suffer from issues like training instability or mode collapse.

Diffusion Models [23] represent a newer class of generative models that have recently

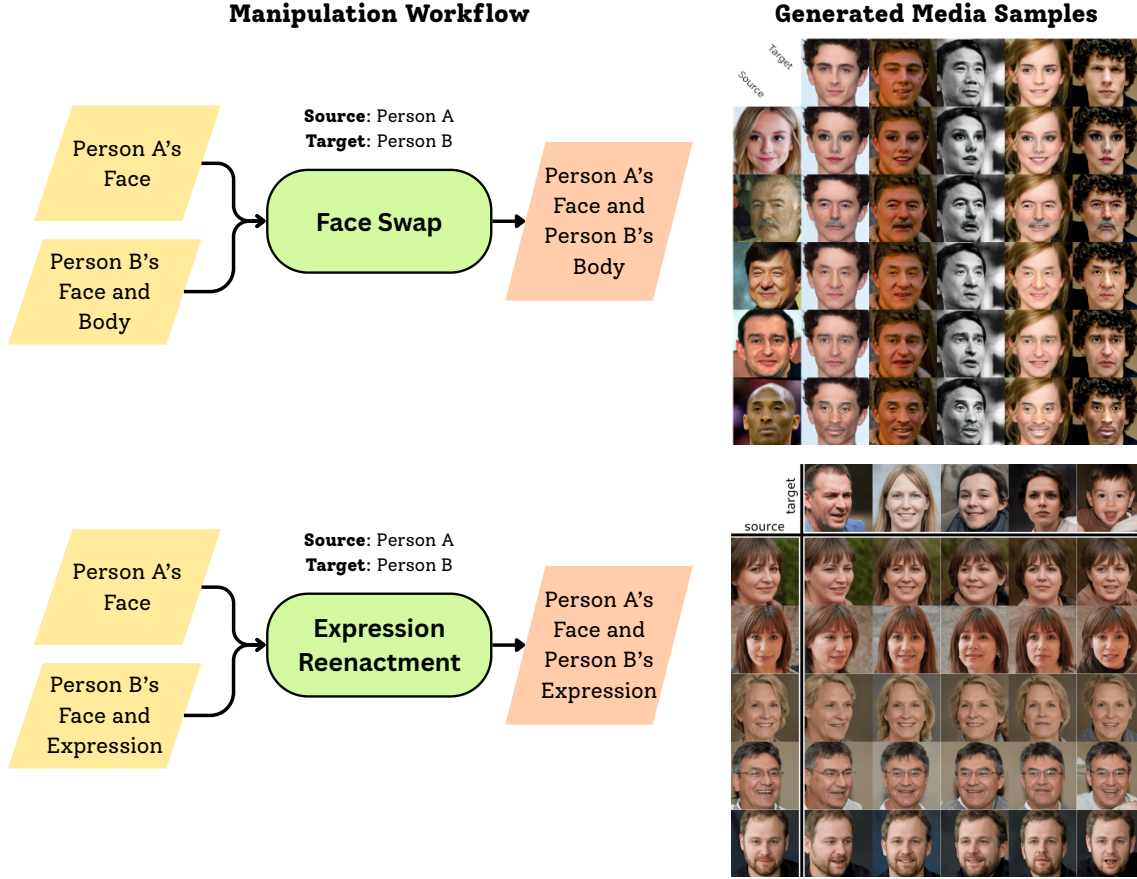


Figure 2.1: Illustration of Deepfake Manipulation Techniques: Face Swapping and Expression Reenactment. The left column outlines the conceptual manipulation workflow, while the right column presents generated media samples. In face swapping, Person A’s facial identity is transferred onto Person B’s body, maintaining the latter’s expressions [9]. In expression reenactment, Person A’s facial identity is retained, but Person B’s expressions are transferred, allowing Person A to mimic Person B’s emotional or speech-related movements [10].

surpassed GANs in synthesis quality. These models operate by iteratively denoising a sample from pure noise to a structured output. Inspired by thermodynamic processes, the generation process is formulated as a gradual reverse diffusion from a noise distribution to a clean image or video. Diffusion-based methods are particularly effective in producing high-resolution, temporally coherent facial animations, and are capable of preserving subtle expression details and identity features. Models like AniFaceDiff [11] and DreamTalk [15] exemplify the success of this approach in talking-head generation. Despite their impressive realism, diffusion models are computationally intensive and typically involve longer inference times compared to GANs.

Neural Radiance Fields (NeRFs) [24] offer a fundamentally different approach by modeling scenes as continuous 3D volumes through coordinate-based neural networks. A NeRF maps spatial coordinates and viewing directions to radiance values, which can be used to synthesize novel views of a scene. In the context of deepfake generation, NeRFs

have enabled the creation of speech-synchronized and photorealistic talking-head videos that include natural head movement and accurate lip synchronization. SyncTalk [4] is a notable example of a NeRF-based model applied to deepfake synthesis. Although NeRFs provide strong 3D consistency and viewpoint generalization, they often require extensive training time and are limited by relatively slow rendering performance.

Together, these generative models define the current landscape of deepfake synthesis, each bringing distinct capabilities and trade-offs that influence the realism, efficiency, and controllability of the generated content. A high-level conceptual comparison of these models is presented in Figure 2.2, illustrating their generation workflows and highlighting the visual differences in how each model produces synthetic faces.

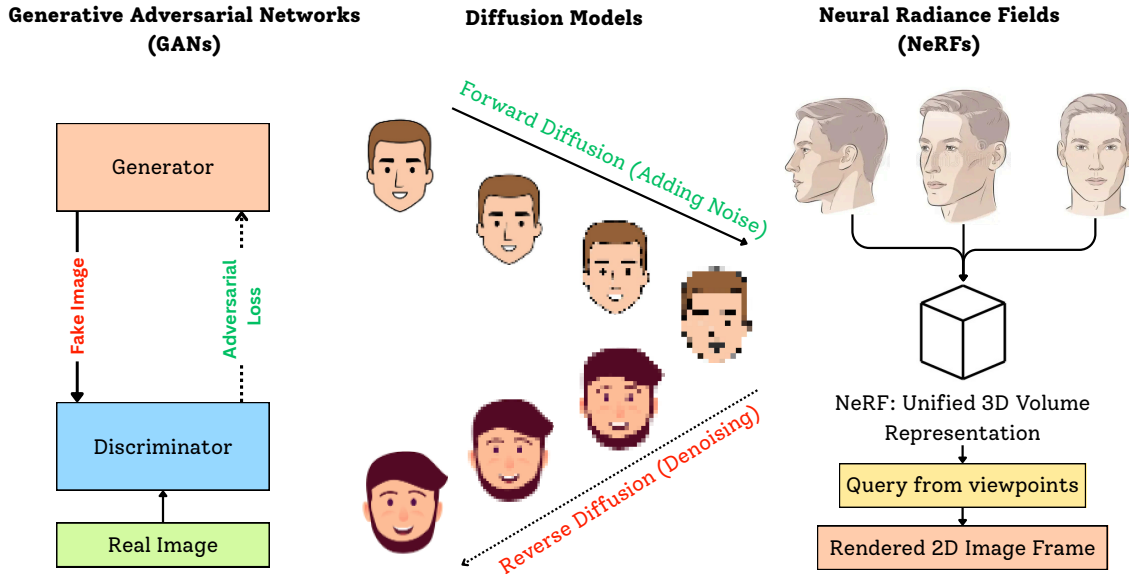


Figure 2.2: Conceptual Comparison of Generative Models Used in Deepfake Synthesis. This diagram illustrates the high-level workflows of three foundational generative models: Generative Adversarial Networks (GANs), Diffusion Models, and Neural Radiance Fields (NeRFs). All models are shown generating facial outputs to highlight their application in deepfake media synthesis. GANs utilize an adversarial training process between a generator and a discriminator. Diffusion models progressively generate images by reversing a noise-adding process. NeRFs represent 3D scenes by learning view-dependent radiance, enabling photorealistic novel-view synthesis from a learned 3D face representation.

2.1.4 State-of-the-Art Deepfake Detection Architectures

Detection methods have evolved from shallow CNNs (e.g., MesoNet [5]) to more advanced models like Xception [30], EfficientNet, Capsule Networks, and Transformer-based detectors. Techniques such as frequency-domain analysis, representation learning (e.g., CORE [31]), and reconstruction-based methods (e.g., RECCE [18]) aim to enhance generalizability across unseen forgeries.

2.1.5 Challenges in Multimodal Deepfake Detection

Modern deepfakes span across modalities—visual (face swaps, reenactment), and audio (voice cloning). Multimodal detection aims to analyze these cross-modal correlations, which is increasingly important as deepfakes become more sophisticated and multimodal in nature.

2.2 Related Works

Traditional generative models have converged on making very realistic synthetic media, effectively making deepfakes difficult to detect. Generative adversarial networks (GANs) were a huge development, but diffusion-based models and Neural Radiance Field (NeRF) have since taken it to an even higher quality and complexity. Using more recent and sophisticated algorithms, we leverage these methods to produce near-lifelike facial expressions, synchronized lip motion, and ‘asymptotic’, as it seems to keep improving head motion as found in AniFaceDiff [11] and DreamTalk [15]. For the detection part, there is some progress on models using Transformer architectures along with contrastive learning. Yet it remains that even the most robust detection frameworks (Xception [30] or CORE [31]), capable of detecting many sophisticated generation techniques, are often unable to match the sophistication of modern generative techniques, leaving a critical gap that calls for future research and invention to keep digital media secure.

Not only the creation, but detection of the deepfake technology has been moving at a rapid pace. There have been a lot of studies on the generation techniques, detection methods, and societal implications of deepfakes from researchers. In this literature review, we compare some of the works that have been written on these aspects and analyze their methodologies, strengths, and weaknesses.

For example, Passos et al. [32], and Wang et al. [33] provide extensive reviews on deep learning-based deepfake detection methods based on CNNs, GANs, and autoencoders. While these papers provide a good overview of the strengths and limitations of each model type to capture deepfake manipulations, in particular in video content, they typically only cover the video-specific detection without fully taking into consideration the intricacies of multimodal content e.g. audio or textual deepfakes. Pei et al. [26] and Juefei-Xu et al. [34] contribute further by benchmarking detection techniques and examining the adversarial ‘arms race’ dynamic between generation and detection. However, it is their work that underlines the ongoing need for detection improvement in the face of growing generative threats, but they note the scalability limitations of addressing the computational requirements of diffusion models and newer synthesis techniques.

Extensively, the societal impacts and ethical dimensions of deepfake technology are

explored. Beyond the impact on specific industries, the discussion of broader implications of deepfakes is discussed by Mubarak et al. [19] and Whittaker et al. [35], with calls for regulatory frameworks and ethical considerations due to issues of privacy, misinformation, and public trust. The studies bridge between technical solutions and social implications, arguing for multi-disciplinary approaches that include legal and ethical aspects. Moreover, targeted studies including Rahman et al. [25] and Waseem et al. [20] discuss some specific types of deepfake (face swapping and video manipulation) in detail but confine their scope to a specific type of manipulation. Heidari et al. [27] provide a systematic review of deep learning-based detection methods according to media type (video, audio, and hybrid) but mainly for visual data and do not include non-deep learning techniques which may be successful in different circumstances.

Lastly, research into human factors in deepfake detection, as shown by Bray et al. [36], presents a unique angle on human-centered approaches, exploring human detection accuracy and intervention techniques. However, the literature as a whole has only limited discussion of the potential synergy between human intuition and machine learning models in identifying deepfakes, with most studies treating human and machine capabilities as separate approaches.

Table 2.1 provides a structured comparison of the existing literature, each study contributing unique insights into deepfake detection and societal impacts. However, several gaps emerge when examining focus areas, methodologies, and practical applicability. Notably, most studies, such as those by Passos et al. [32] and Heidari et al. [27], rely heavily on secondary data for testing detection models. This reliance on secondary datasets limits these studies' real-world applicability, as they are not empirically validated in varied, uncontrolled environments where new generation techniques and novel types of deepfakes may arise.

2.2.1 Summary of Gaps and Research Questions

Through an extensive review of prior work, the following research gaps have been identified—each of which is directly explored in this thesis:

- *Generalizability Gap*: Many detectors fail when applied to content generated by unseen or newer manipulation methods, such as diffusion-based models. This research evaluates detection performance across diverse generative techniques.
- *Human-AI Comparative Gap*: Few studies systematically compare algorithmic detection with human perception to understand their respective strengths. This thesis includes a large-scale human study alongside model benchmarking.
- *Modality Gap*: Most existing detection models focus exclusively on visual features, overlooking the potential contribution of audio signals. This thesis investigates

whether incorporating audio enhances detection performance.

- *Fusion Strategy Gap*: While multimodal detection has gained traction, limited work rigorously compares different fusion approaches. This thesis experimentally evaluates early, late, and cross-modal attention fusion strategies using pretrained models.

These gaps lead to the following research questions explored in this work:

- How do detection models perform across manipulation types, particularly on newer diffusion-based content?
- How does human performance compare to state-of-the-art detectors in detecting deepfakes under varying conditions?
- Does the addition of audio modality significantly enhance deepfake detection performance over visual-only models?
- Which multimodal fusion strategy—early fusion, late fusion, or cross-modal attention—yields the most effective performance?

These questions form the foundation for the more targeted gap discussion presented in Section 2.3.

Table 2.1: Comparative Summary of Existing Review Articles on Deepfake Generation vs Detection: Focus Areas, Methodologies, and Research Contributions

Paper Title	Focus	Methodologies	Strengths	Limitations	Key Contributions	Generative Models Used	Experiment Methodology	Human Survey	Detection Emphasis	Evaluation Metrics
Deepfake Detection Using Deep Learning Methods (WIREs 2024) [27]	Detection Review	Systematic review on deep learning methods by modality	Comprehensive categorization for various media	Focus on DL, limited non-DL discussion	Structured overview of DL-based detection	GAN-based	Secondary data	No	DL techniques	Precision, recall, F1
Deepfake Detection: A Comprehensive Survey (ACM 2024) [33]	Reliability	Focus on model reliability, transferability, and robustness	Insights on reliability and generalization	Limited real-time applicability	Reliability perspective for real-world detection	GAN-based	Self-conducted experiments	No	Generalization	F1, AUC
A Review of Deep Learning Approaches for Deepfake Detection (Expert Systems 2024) [32]	Deepfake Detection	Review of CNNs, GANs, and autoencoders for detection	In-depth analysis of GAN manipulations	Focused on video, limited multi-modal coverage	Systematic deep learning approach for detection	GAN-based	Secondary data	No	CNNs and autoencoders	Precision, recall
A Survey on the Detection and Impacts of Deepfakes (IEEE Access 2023) [19]	Societal Impacts	Analysis of deepfake impacts across visual, audio, and textual domains	Holistic view of societal impact	Limited technical detection focus	Combines tech insights with societal impact, calls for legislation	GAN-based	Secondary data	No	Multimodal (visual, audio, text)	Thematic analysis
DeepFake on Face and Expression Swap (IEEE 2023) [20]	Face Swap Review	Reviews GAN and autoencoder-based face swap techniques	Detailed analysis of face swap, future detection directions	Focus only on facial manipulation	Insights into facial manipulation techniques	GAN-based	Secondary data	No	GAN-based methods	Mean average precision
Mapping the Deepfake Landscape (Technovation 2023) [35]	Multidisciplinary Review	Multidisciplinary view of deepfake tech, legal, and social aspects	Cross-disciplinary impact, innovation	Lacks technical detection detail	Calls for collaborative solutions in tech, law, and social science	GAN-based	Secondary data	No	Multidisciplinary impact	Thematic analysis
Testing Human Ability to Detect Deepfake Images (JCS 2023) [36]	Human Detection	Experimental study on human accuracy in detecting deepfakes	Provides empirical evidence on human detection accuracy	Limited technical detection focus	Challenges in human detection, suggests improved automated tools	GAN-based	Self-conducted survey	Yes	Human perceptual cues	Confidence score, accuracy
Countering Malicious DeepFakes (IJCV 2022) [34]	Adversarial Deepfake Landscape	Focus on the adversarial relationship between generation and detection	Highlights arms race between generation and detection	Limited solution coverage, focus on dynamics	Sets research agenda for adaptive detection methods	GAN and Diffusion	Secondary data	No	GAN and adversarial dynamics	Accuracy, AUC
A Qualitative Survey on Deepfake Video Creation and Detection (AJEIT 2022) [25]	Creation and Detection Review	Comparison of traditional and DL-based methods for video creation and detection	Highlights evolution of deepfakes	Lacks quantitative benchmarks	Broadened understanding of creation and detection techniques	GAN-based	Secondary data	No	Conventional and DL-based methods	Case study analysis

2.3 Gap Analysis

A significant gap also exists in the exploration of human involvement in deepfake detection. Only Bray et al. [36] incorporated a human survey, assessing the capabilities and limitations of human detection accuracy, but most studies exclude any examination of how human discernment might complement machine detection methods. This oversight misses an essential aspect of detection: the potential synergy between human intuition and machine algorithms, which could enhance the accuracy and adaptability of deepfake detection in complex real-world settings. Additionally, current research lacks adaptability to emerging generative techniques. Most studies are grounded in traditional deep learning models (primarily GAN-based) and single-modality applications, which restricts their scope as deepfake content increasingly spans multiple modalities (visual, audio, and textual). While prior work has explored multimodal detection [37, 38, 39, 40], many models remain predominantly vision-focused and do not fully leverage audio information. This thesis examines whether integrating audio improves performance and generalizability across manipulation types. Further, existing benchmarks remain static, limiting evaluations to conventional generative techniques rather than testing against cutting-edge models like NeRF and diffusion models, which produce sophisticated and highly realistic content.

To address the limitations identified in the existing literature, this thesis proposed targeted contributions aimed at advancing the field of deepfake detection. The key research gaps and the corresponding approaches taken to address them are outlined in Table 2.2. This table highlights the study’s emphasis on improving real-world applicability, incorporating insights from human observation, and developing a comprehensive multimodal detection framework.

Table 2.2: Addressing Key Gaps in Existing Deepfake Literature: Methodological Enhancements and Focus Areas

Literature Gap	How This Work Addressed the Gap	Focus
Reliance on Secondary Data: Most studies relied on secondary data for validation, limiting real-world applicability and adaptability to emerging generative methods.	This work included self-conducted experiments and empirical testing using recent generative techniques (e.g., NeRF and diffusion models) in varied, real-world contexts.	Conducted primary, real-world experiments
Limited Insights from Human Observation: Few studies explored how humans perceive and detect deepfakes, missing insights into human detection strategies.	This work conducted a human survey to gather detailed insights into human observation of deepfakes, offering a nuanced understanding of perception and detection strategies.	Provided insights on human perception and detection thinking
Lack of Empirical Evaluation of Multimodal Detection: While some models use multimodal inputs, few studies systematically evaluate whether audio modality truly enhances detection performance.	This work investigates the effectiveness of incorporating audio modality through an experimental framework using pretrained models and fusion strategies, offering empirical evidence to guide future multimodal research.	Explored and validated audio-visual integration for future detection systems

As shown in Table 2.2, these contributions strengthened the robustness and adaptability of deepfake detection. The detection model was enhanced in practical relevance

through real-world experiments and validated against cutting-edge generative techniques, ensuring adaptability to fast-evolving deepfake synthesis methods. The incorporation of human observational insights offered an additional perspective on strengths and limitations in deepfake recognition, informing perceptual-based detection strategies. Lastly, this work experimentally investigated whether the inclusion of audio modality enhances detection performance, offering empirical guidance for future research in designing effective multimodal deepfake detection systems.

This work laid the foundation for future deepfake detection research by introducing an adaptable and comprehensive detection framework that leveraged empirical validation, human-centered insights, and multimodal analysis. These contributions addressed critical gaps in the literature and provided a more resilient and holistic defense against the growing challenges posed by deepfake technologies.

2.4 Summary

This chapter established the necessary background for understanding the domain of deepfake generation and detection. We introduced fundamental concepts such as GANs, diffusion models, and NeRFs, and reviewed a wide range of detection architectures from shallow CNNs to Transformer-based models. Through a comprehensive literature review and a detailed gap analysis, we identified key limitations in the current landscape: a lack of real-world validation, underexplored human-machine synergy, and insufficient empirical studies on multimodal detection frameworks. These insights directly motivate the contributions of this thesis, which addresses these challenges by conducting real-world experiments, analyzing human perceptual performance, and empirically evaluating the impact of audio-visual integration—ultimately guiding future research on human-aligned, multimodal deepfake detection systems.

Chapter 3

Methodology and Experimental Design

This chapter outlines the methodological approach adopted for evaluating deepfake detection techniques. It covers the requirement analysis, detailed methodology, and the empirical evaluation of various detection models and human-centered assessments.

3.1 Requirement Analysis

In this section, we focus on the requirements essential for conducting a comprehensive empirical evaluation of deepfake detection and generation techniques. The goal is to assess the performance of various detection models against modern deepfake synthesis methods, including GAN-based and diffusion-based models. Given the rapid evolution of generative techniques, it is imperative to evaluate the robustness of current detection algorithms and identify areas where improvements are needed.

The primary requirements for this study include:

- **Diverse Dataset:** The evaluation requires a rich and diverse dataset comprising both real and synthetic videos. The synthetic data should be generated using state-of-the-art techniques, including GANs, diffusion models, and NeRF-based approaches, to reflect the diversity of deepfake methods.
- **Detection Algorithms:** Selection of robust deepfake detection models ranging from classical CNN-based architectures (like Xception and MesoNet) to advanced models incorporating Transformers and contrastive learning techniques. The chosen algorithms should represent both established and cutting-edge detection methods.
- **Human Evaluation:** Conducting a human-centered study to gauge human ability to detect deepfakes. This step is essential to compare algorithmic detection accuracy against human perceptual capabilities, providing insights into the strengths and limitations of both approaches.

- **Benchmarking Framework:** A comprehensive framework for benchmarking both automated and human detection results, focusing on metrics such as AUC, Average Precision (AP), Precision, Recall, Sensitivity Index (d-prime), and C-Decision Criterion (Bias). This framework must accommodate various video conditions, including high and low resolutions. These metrics are essential in the context of deepfake detection to ensure a balanced evaluation of detection accuracy, sensitivity, and bias. AUC (Area Under the Curve) measures the overall ability of the system to distinguish between real and fake videos across all decision thresholds, providing a threshold-independent evaluation of discriminative performance. Average Precision (AP) summarizes the precision-recall trade-off by calculating the average of precision values at different recall levels, which is especially valuable in imbalanced datasets where fake videos are rare. Precision reflects how many of the detected deepfakes are actually fake, reducing false accusations. Recall (or sensitivity) measures how many actual deepfakes were correctly identified, helping to ensure that manipulative content does not go undetected. Sensitivity Index (d-prime) captures how well a system can separate signal (deepfakes) from noise (real videos), offering insight into its detection sensitivity regardless of decision threshold. C-Decision Criterion (Bias) quantifies the system’s tendency to lean toward false positives or false negatives—critical in deepfake detection where over-flagging can erode trust in real content and under-flagging can let harmful media spread. Together, these metrics provide a comprehensive view of detection performance, guiding the development of systems that are both accurate and trustworthy under varying video conditions.
- **Multimodal Integration:** Investigating whether incorporating audio features alongside visual data can enhance deepfake detection performance. This requirement addresses the growing need to understand the contribution of cross-modal signals in detecting increasingly sophisticated audio-visual deepfakes.

3.2 Experimental Design

To fulfill the requirements outlined in Section 3.1, this section presents how each research objective was implemented through a structured experimental framework. The design spans four key components—data diversity, model benchmarking, human evaluation, and multimodal analysis—each of which is mapped directly to the functional requirements previously specified.

First, in response to the requirement for a diverse and representative dataset, we selected a set of modern deepfake generation models that reflect the current state of synthesis techniques. These include GAN-based (e.g., FaceVid, FaceFusion), diffusion-based (e.g., AniFaceDiff, DreamTalk), and NeRF-based (e.g., SyncTalk) approaches. Rather than reproducing these models, we focused on reviewing their mechanisms, highlighting their conditioning strategies and realism characteristics, and justifying their inclusion for bench-

marking purposes. This selection rationale and model overview are provided in Section 3.2.1.

To address the need for rigorous evaluation of deepfake detection methods, we implemented and compared ten state-of-the-art detection models, including MesoNet, Xception, CORE, RECCE, and UCF. These models were chosen to represent a diverse range of architectural paradigms, such as CNNs, capsule networks, contrastive learning, and attention-based mechanisms. Each model was trained and evaluated using standardized metrics (AUC, AP, precision, recall, d-prime, and C) across multiple synthesis methods. The details of this consolidation and comparison are presented in Section 3.2.2.

To fulfill the requirement for human-level performance benchmarking, we conducted a large-scale user study involving 271 participants who evaluated a carefully curated set of real and deepfake videos generated using various synthesis methods and resolutions. Participants’ responses were collected through a Likert-style interface and later transformed into probabilistic scores to enable direct comparison with model predictions. The design of the survey—including participant demographics, video selection, and response mechanisms—is described in Section 3.2.3, while the analysis of participants’ detection performance and perceptual tendencies is presented in Chapter 4.

Lastly, to explore the requirement of multimodal detection, we designed a detection framework that integrates audio and visual features using pretrained models—TimeSformer for video and Wav2Vec 2.0 for audio. Importantly, we do not propose a novel state-of-the-art multimodal detection model. Rather, the goal of this component is to investigate whether incorporating audio improves deepfake detection performance compared to unimodal baselines. To this end, we implemented three fusion strategies—early fusion, late fusion, and cross-modal attention—and evaluated them on the FakeAVCeleb and AV-Deepfake1M datasets. This experimental setup and its rationale are described in Section 3.2.4.

Together, these four components operationalize the methodological requirements established in Section 3.1, forming a comprehensive framework to evaluate both algorithmic and human deepfake detection performance under diverse and realistic conditions.

3.2.1 Data Preparation: Deepfake Synthesis

Deepfake synthesis has seen rapid progress in recent years, driven by advancements in generative modeling and deep learning. Modern approaches focus on enhancing the visual quality, identity preservation, and expression fidelity of synthesized faces, especially in the context of face reenactment. This section provides an overview of several state-of-the-art deepfake synthesis models, highlighting their unique architectural designs, conditioning mechanisms, and performance on benchmark datasets such as VoxCeleb.

AniFaceDiff: AniFaceDiff [11] is a face reenactment method built upon the Stable Diffusion framework. It introduces a novel conditioning module designed to enable high-fidelity facial animation. The primary innovations include an improved 2D facial snapshot

conditioning mechanism that aligns facial shapes to mitigate the transfer of irrelevant identity information, and an expression adapter that effectively preserves nuanced expression details. This approach maintains a balance between the preservation of pose and expression fidelity from the driving video and the retention of identity and fine-grained details from the source image. The method was evaluated on the VoxCeleb dataset, where it demonstrated state-of-the-art performance in face reenactment, particularly in cross-identity scenarios.

Thin-Plate Spline Motion Model (TPSMM): TPSMM [2] leverages Thin-Plate Spline Motion Estimation to generate a more flexible optical flow, enabling the warping of feature maps from the source image to align with the feature domain of the driving image. This flexibility allows for better handling of large pose variations. The model incorporates Multi-Resolution Occlusion Masks to more realistically restore missing regions through improved feature fusion, ensuring the visual coherence and realism of the generated animation. Additionally, carefully designed loss functions promote a clear division of responsibilities among network components, further enhancing the quality and realism of the output images. The VoxCeleb dataset was used for both training and evaluation.

FaceVid: FaceVid [1] presents a neural talking-head video synthesis framework that utilizes keypoint representations to decompose identity-specific and motion-related information in an unsupervised manner. As one of the earlier works in the field of face reenactment, it adopts a GAN-based architecture and provides foundational insights for subsequent advancements in neural talking-head generation.

HyperReenact: HyperReenact [14] leverages the photorealistic generation capabilities and disentangled representations of a pretrained StyleGAN2 generator. The method first inverts real images into the latent space of StyleGAN2 and then employs a hypernetwork to refine the identity features of the source image while re-targeting the facial pose. Operating in a one-shot setting—where only a single source frame is used—HyperReenactment enables cross-subject reenactment without requiring subject-specific fine-tuning. The method was trained and evaluated on the VoxCeleb1 and VoxCeleb2 datasets.

FaceFusion: FaceFusion [12] is a continuously updated open-source face-swapping tool that also incorporates frame enhancement capabilities for both images and videos. It originated as a fork of the popular face-swapping application Roop, which gained popularity for its ease of use and integration with Stable Diffusion-based extensions. Building upon Roop’s foundation, FaceFusion [12] introduces a suite of new features and performance improvements. It utilizes multiple models such as BlendSwap, InSwapper, and SimSwap. However, details regarding the training procedures and implementation methods remain unclear, as they have not been formally published.

FADM: Face Animation with an Attribute-Guided Diffusion Model (FADM) [13] employs diffusion models to generate photorealistic talking-head videos. To address the uncontrolled synthesis effects commonly encountered in diffusion-based generation, the authors introduce the Attribute-Guided Conditioning Network (AGCN). AGCN adaptively fuses coarse animation features with 3D face reconstruction results, allowing the

diffusion model to incorporate both appearance and motion cues effectively. FADM corrects unnatural artifacts and distortions, while iteratively refining facial details with high fidelity. The model was trained on VoxCeleb1, VoxCeleb2, and CelebA datasets.

VASA-1: VASA-1 [3] excels in generating highly synchronized lip movements in alignment with the input audio while capturing subtle facial expressions and natural head motions that enhance realism. The core innovations of VASA include a unified model for synthesizing facial dynamics and head movements within a learned face latent space, as well as the construction of a disentangled and expressive latent space trained on video data. These design choices allow the system to produce high-quality talking-head videos with realistic expression and motion. Moreover, the model supports real-time generation of 512×512 resolution videos at up to 40 frames per second with minimal latency. As the model was not released publicly, videos were obtained from the authors’ project webpage.

DreamTalk: DreamTalk [15] is a diffusion-based framework for synthesizing emotionally expressive talking-head videos. It comprises three primary components: a denoising network based on diffusion for consistent and high-quality audio-driven facial motion generation across varying emotional states; a style-aware lip-sync module that improves lip synchronization while preserving the intensity of emotional expressions; and a style predictor that learns personalized emotional styles directly from audio input, thus eliminating the need for additional emotion references.

Diffused Heads: Diffused Heads [16] proposes an autoregressive diffusion-based model for talking-head video generation that addresses longstanding challenges related to the synthesis of natural head movements and facial expressions. The model requires only a single identity image and an audio clip to generate highly expressive videos. It is capable of hallucinating realistic facial behaviors such as blinks and head tilts, while maintaining the original background. Evaluation on two benchmark datasets demonstrates that DiffusedHeads achieves state-of-the-art results in terms of expressiveness and temporal smoothness.

SyncTalk: SyncTalk [4] is the first NeRF-based approach designed to synthesize highly synchronized, speech-driven talking-head videos with photorealistic quality. It introduces a Face-Sync Controller to ensure precise alignment between lip movements and speech, along with a 3D facial blendshape model for accurately capturing facial expressions. Furthermore, a Head-Sync Stabilizer refines head movements for enhanced naturalism, while the Portrait-Sync Generator reconstructs hair details and seamlessly blends the generated head with the torso region.

To emphasize current progress in the field, we group videos generated by SyncTalk [4], DreamTalk [15], Diffused Heads [16], and VASA-1 [3] under the category *SOTA* (State-of-the-Art), as these models consistently demonstrate superior visual realism and temporal coherence. Notably, the codebases for these models were not publicly released, primarily due to ethical considerations surrounding potential misuse. As a result, the corresponding videos used in this study were obtained directly from the authors’ official project repositories or supplementary materials. Representative sample frames from the selected

synthesis methods are illustrated in Figure 3.1, and a summary of their key characteristics is presented in Table 3.1.

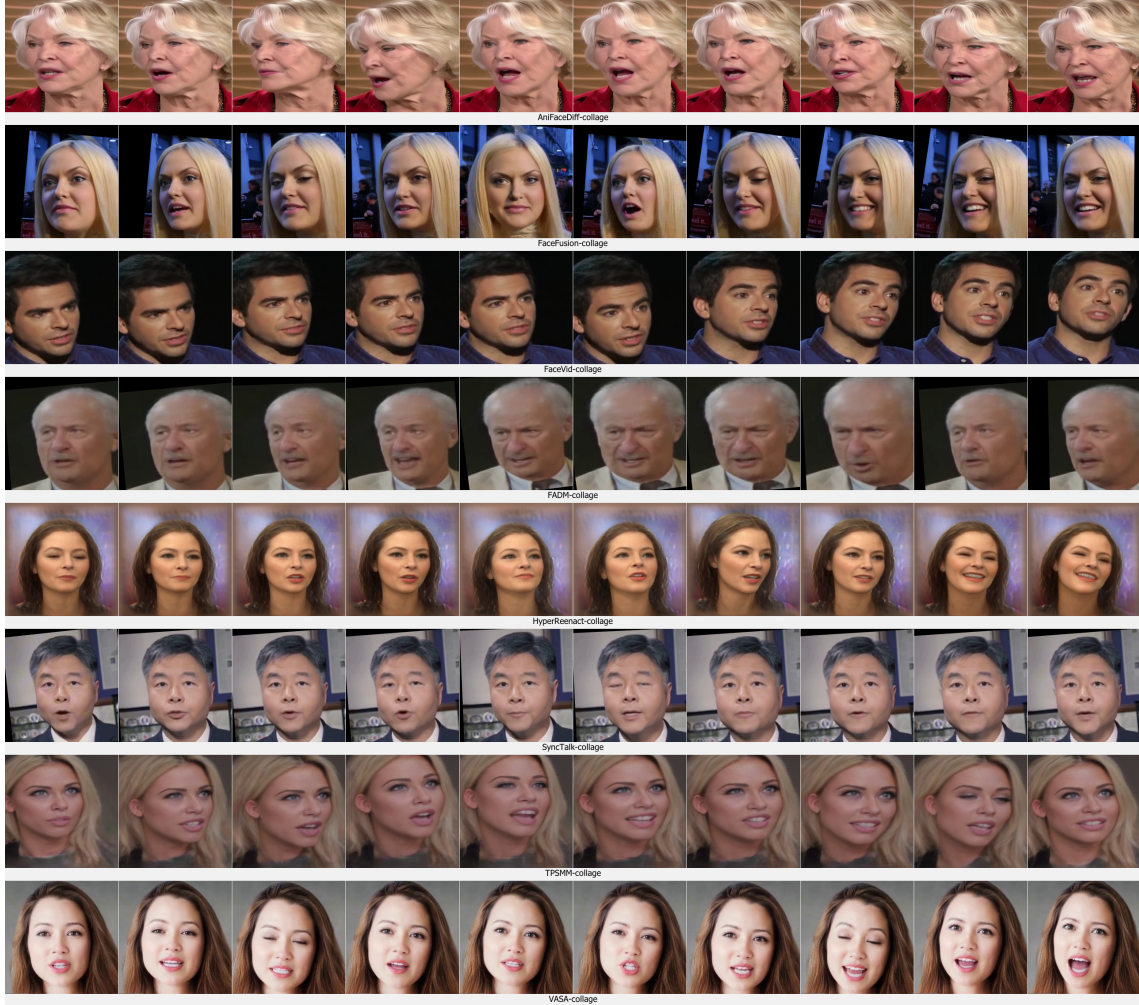


Figure 3.1: Sequence of frames from videos generated using the deepfake synthesis methods discussed in this paper. The figure showcases high-quality and realistic deepfake videos, demonstrating realistic facial movements that are challenging to detect with casual inspection. Methods depicted in the figure are respectively: AniFaceDiff [11], FaceFusion [12], FaceVid [1], FADM [13], HyperReenact [14], SyncTalk [4], TPSMM [2], and VASA-1 [3].

3.2.2 Deepfake Detection Algorithm Consolidation

With the rapid advancement of deepfake generation techniques, the development of reliable and generalizable deepfake detection methods has become increasingly critical. This section provides an overview of prominent deepfake detection models, highlighting their underlying architectures, key innovations, and performance on benchmark datasets. These methods aim to identify subtle artifacts and inconsistencies introduced during manipulation, with approaches ranging from early CNN-based frameworks to recent models lever-

Table 3.1: Comprehensive summary of various deepfake generation techniques, detailing the method, datasets and number of images/videos used, and key highlights for each approach.

Name	Published in	Method	Dataset Used	Type	Number of Images/Videos	Highlights
FaceVid [1]	2021	GAN	VoxCeleb2, TalkingHead-1KH	Image-to-video	1M (VoxCeleb2), 180K (TalkingHead-1KH)	Generates realistic talking-heads from a single image, separates appearance and motion using keypoints, high fidelity through warping-based generation.
FaceFusion [12]	2022	GAN	Based on Wav2Lip, InsightFace	Video-to-video	N/A	Combines face swapping and lip-syncing in a modular real-time system, preserves identity and expression using multi-scale blending.
TPSMM [2]	2022	Thin-plate-spline	VoxCeleb1, CelebV, MUG	Image-to-video	100K videos	Learns non-rigid deformations via TPS, improves generalization in one-shot scenarios, and enhances motion realism using dense warping.
HyperReenact [14]	2023	StyleGAN-2	VoxCeleb1, HDTF, CelebV-HQ	Video-to-video	100K videos	Employs hypernetworks and 3DMM controls for identity-preserving and expressive reenactment, robust to pose and lighting variation.
FADM [13]	2023	Diffusion	VoxCeleb2, CelebV, HDTF	Image-to-video	100K+ (VoxCeleb), 1M (VoxCeleb2), 200K (CelebA)	Uses attention-based decomposition of expression and pose, enables accurate and high-quality face animation across multiple datasets.
DreamTalk [15]	2023	Diffusion	VoxCeleb2, HDTF, MEAD	Audio-to-video	280K (MEAD, HDTF), 1M (VoxCeleb2)	Two-stage diffusion pipeline with multi-modal conditioning, excels in audio-lip-text alignment and photorealistic facial generation.
AniFaceDiff [11]	2024	Diffusion	VoxCeleb2, TEDXPeople, MUG, CelebV-HQ	Audio-to-video	1M videos (VoxCeleb2)	Dual-path diffusion transformer for expressive, identity-consistent talking-heads, supports zero-shot generation with high realism.
VASA-1 [3]	2024	Diffusion Transformers	VASA, LRW, LRS3, CelebV	Audio-to-video	1M (VoxCeleb2), 22K (VASA)	Offers real-time and zero-shot talking-head generation with gaze, emotion control, driven by latent diffusion & transformer models.
Diffused Heads [16]	2024	Diffusion	VoxCeleb2, LRW, HDTF, MEAD	Audio-to-video	100K+ (LRW)	Uses 3D face priors with dual-branch diffusion to decouple geometry and texture, enabling high control over expressions and identity.
SyncTalk [4]	2024	NeRF	VoxCeleb2, LRS2, CelebV-HQ	Audio-to-video	1M videos, 8K+ frames/video	Introduces sync-aware loss and multimodal attention, achieves superior lip sync and visual quality, suitable for real-time generation.

aging attention mechanisms, representation learning, and disentangled feature extraction.

MesoNet: Using deep learning, MesoNet [5] is an early method in detecting deepfake videos, together with two networks named Meso-4 and MesoInception-4. The reason they achieve such high results is by carefully studying a face’s features, including mouth and eyes which are vital for noticing changes. The technique is best for managing issues resulting from compression of videos and images that are blurry because forged.

Its special network structure helps MesoNet notice and highlight the types of abnormalities that are common in altered videos. It relies on mesoscopic techniques which focus on both the detail of the face and its environment, to tell true faces from false ones. Furthermore, training MesoNet on a diverse data set increases its ability to resist different types of Deepfake techniques and generalize to new data.

Xception: XceptionNet [6] was created by altering a CNN originally trained on ImageNet and adding two end outputs specific for detecting deepfakes. First, the network trains itself for 3 epochs using the pre-set parameters, then for 15 more epochs it trains for better validation results. It has demonstrated superior abilities compared to other methods at detecting changed images, especially when it comes to faces.

Xception [6] uses depthwise separable convolutions to ensure the model runs faster and more accurately, even with less numbers of parameters. In order to detect deepfakes, it scans both space and time in video frames for inconsistencies between the real and fake videos. The structure of the model allows it to notice the details and out-of-the-ordinary aspects common in deepfakes.

Capsule: To detect deepfakes, capsule [41] networks are used because they can recognize how parts of objects interact in a three-dimensional way and improve the method of spotting fake images. To decide whether an image or video is real or fake, the proposed model processes them with VGG-19 and uses the extracted features in its main capsules to classify the contents. Taking this method into account, the system can deal with forged scenarios such as replay attacks and computer-made content, more efficiently than other approaches.

Traditional CNNs sometimes miss how features are structured relative to one another, but capsule networks charge focus on these details and hence, aid in spotting deepfake images. FEATURE detectors emit vectors inside capsules, allowing the model to identify and understand the position of facial features and their actions. As a result, capsule networks can identify the difference between real and fake content, even if the deepfake is challenging to spot.

EfficientNet: Even though it was not created for detecting deepfakes, EfficientNet [17] is still suitable for the task since it is highly accurate and efficient. While achieving leading results on different datasets, the model uses much less memory than other convolutional models. Because EfficientNet is capable of transfer learning, it can be improved for spotting deepfakes.

It does this by using a technique that scales three aspects of the network together to achieve more accurate results even with fewer computations. It has the ability to discover

small details that can tell whether an image or video is genuine or fake. Many researchers prefer EfficientNet in their deepfake detection frameworks, as it is shown to achieve better results and higher precision and recall than previous models.

FFD: FFD [42] has been created to help in both developing and evaluating deepfake detection models. It holds a vast number of video clips made through face swaps using technologies like Face2Face, DeepFakes, FaceSwap and NeuralTextures. The dataset is available in two versions: an uncompressed version and one that is compressed up to half its size.

FFD [42] assists in deepfake detection by providing a common framework for all tests. It relies on a CNN model called XceptionNet to perform baseline detection, as it effectively learns features using depthwise separable convolutions. The models learn to find inconsistencies in certain areas that appear when frames are tampered with. Besides, the report notes that a model’s detection ability declines when it is compressed, making it important to train on various and realistic images to avoid issues with generalization.

SRM: SRM (Spatial Rich Model) [7] is a method for deepfake detection that extracts more information from image noise at high frequencies. It recognizes changes in the noise due to modification because they tend to create specific features in the disguised image. The inclusion of different noise features in SRM allows forgery detection models to be stronger against changes made before forgery detection.

It analyzes the details in images based on their spatial features, mostly on residual noise, to determine if anything has been changed. It has a module that identifies these slight discrepancies that ordinary methods might miss. Paying particular attention to different areas in the data such as with residual guided spatial attention and dual cross-modality attention, plays a big role in making SRM more effective.

RECCE: RECCE [18] is a new method that seeks to identify the main differences between genuine and fake faces. Reconstruction learning, multi-scale graph reasoning and reconstruction guided attention are the three main components and they are designed to make the model capable of finding unknown ways that a face might be changed by just focusing on how real faces are differently organized. The framework makes use of joint optimization to improve accuracy and resistance against several types of disturbances.

This method (RECCE) uses a reconstruction learning process and builds a model designed to identify if images are real or if they have been edited. Thanks to the use of multi-scale graph reasoning, it is able to detect details at different sizes and notice relationships among them. Using the reconstruction guided attention makes it easier for the model to find and highlight the parts in images prone to having inconsistencies, providing protection from wide-ranging criteria of deepfake methods.

CORE: CORE [31] detects deepfakes by representing an input in different ways using data augmentation and requiring the representations to be consistent using a Consistency Loss. By looking at important features, rather than modified content, this approach achieves better results when compared to modern techniques. CORE can be easily combined with various other models and has demonstrated improved performance in recog-

nizing items in many datasets.

With a dual-branch design, CORE has one branch that handles the original input and the other that applies tweaks, so that consistent outputs are provided for every real image. By emphasizing features that are not affected by common changes, the method allows the detection of real content from fake content. Directed representations provided by CORE allow the model to identify and defeat various types of deepfakes.

UCF: UCF (Uncovering Common Features) [43] tackles problems related to being able to use a model for fresh formats of deepfakes. It separates parts of the image using a disentanglement approach and trains AI to spot differences between the source image and image with forgery using multi-task learning. It outperforms current popular methods in applying to unseen data sets.

The method UCF explores shared aspects across several deepfake models, allowing it to be more effective against different types of false images. With a multi-task framework, the model can learn to distinguish both the differences and similarities between real and deepfake images. CelebDF and DFDC results reveal that the model succeeds in outperforming other detectors, showing that it can work well in a wide range of cases.

As demonstrated in Table 3.2, a wide range of deepfake detection techniques have been proposed in recent years, each utilizing different datasets and methodologies to address the growing challenge of face forgery detection. The table summarizes key characteristics of these methods, such as the type of network architecture employed, the scale of the datasets used for training, and the specific innovations introduced to enhance detection accuracy.

For instance, the MesoNet model, which builds upon the Meso-4 architecture, incorporates inception modules and skip connections to achieve a high detection rate, particularly for Deepfake videos. On the other hand, more recent approaches like CORE and UCF emphasize consistency learning and disentanglement frameworks to improve generalization across different datasets. Notably, EfficientNet-B4’s use of compound scaling highlights a trend towards optimizing models for both accuracy and computational efficiency, a crucial factor given the large scale of datasets like ImageNet.

Overall, Table 3.2 serves as a valuable resource for understanding the progression of deepfake detection techniques, from early methods focusing on specific types of forgeries to more sophisticated approaches capable of handling a broader range of manipulations. This comparative analysis underscores the importance of continuous innovation in the field to keep pace with the evolving nature of deepfake generation technologies.

3.2.3 Human Level Detection

This study assessed participants’ ability to detect deepfakes across various video types and resolutions using an anonymous, IRB-approved survey, with informed consent obtained from all participants. To evaluate human-level deepfake detection capabilities and attitudes toward AI-generated media, we designed a comprehensive questionnaire hosted

Table 3.2: Summary of different deepfake detection methods, including their publication details, datasets, techniques, model types, and key highlights.

Name	Publishing	Dataset trained on	Method	Type	Number of images/videos	Highlights
MesoNet [5]	WIFS 2018	Deepfake, Face2Face [44]	Incorporates modified inception modules with 3×3 dilated conv. and 1×1 conv and skip-connections, builds upon the Meso-4 architecture	Naive	1000+ videos, DeepFake & Face2Face	Comprises four layers of convolution and pooling for feature extraction, Achieves over 98% detection rate for Deepfake videos, Contains 27,977 trainable parameters for efficient processing, Employs ReLU activation, Batch Normalization, and Dropout for improved generalization, Specifically designed to detect face tampering in videos
Xception [6]	ICCV 2019	ImageNet [45], fine tuned on FaceForensics++ [6]	CNN architecture with separable convolutions and residual connections, pre-trained on the ImageNet [45] dataset	Naive	ImageNet: 14m images, FaceForensics++: 1000 real, 4000 fake videos	Outperforms other architectures in face forgery detection, Utilizes pre-trained weights from ImageNet for better generalization, Final layer modified for two outputs specific to forgery detection, Pre-trained for 3 epochs, then trained for 15 more epochs, Maintains performance on compressed videos
Capsule [41]	ICASSP 2019	FaceForensics [46]	Capsule networks, Dynamic routing algorithm (VGG-19 network) [47] for feature extraction	Spatial	~1000 videos	Lower half total error rate than state-of-the-art methods, Improved accuracy using capsule network with VGG-19 features, Perfect accuracy in distinguishing CGI from photographic images, Robustness against adversarial attacks via training with random noise, Validation through comprehensive experiments on forgery detection
Efficient Net-B4 [17]	ICML 2019	ImageNet [45]	Compound scaling for uniform adjustment of network dimensions, EfficientNet, a mobile-size baseline network optimized for accuracy and FLOPS	Naive	~1.2 million images	Introduces a new family of models that optimize accuracy and efficiency through compound scaling; EfficientNet Architecture, Achieves top results on various benchmarks, including ImageNet, with fewer parameters, Proposes a systematic approach to scaling networks in width, depth, and resolution, Demonstrates effectiveness in transfer learning across multiple datasets and tasks, Significantly reduces computational costs while maintaining high accuracy
FFD [42]	CVPR 2020	Diverse Fake Face Dataset [42]	CNN-based network with an attention mechanism, XceptionNet for depthwise separable convolutions and residual connections	Spatial	2.6 million images	Enhances localization of manipulated regions for improved detection accuracy, Achieves AUC of 99.64 with low EER values using XceptionNet, Tested against benchmarks, outperforming existing methods in facial forgery detection, Utilizes both XceptionNet and VGG16 for comprehensive performance analysis
SRM [7]	CVPR 2021	FaceForensics++ [6]	Two-stream network architecture, residual guided spatial and dual cross-modality attention modules, Xception backbone	Freq.	1000 real, 4000 fake videos	Two-stream network for face forgery detection, Residual guided spatial and dual cross-modality attention modules, Outperforms existing methods on multiple datasets, Validates effectiveness through extensive testing, Acknowledges backing from research programs and funding bodies
RECCE [18]	CVPR 2022	FaceForensics++, Celeb-DF, WildDeepfake, DFDC	Reconstruction learning, multi-scale graph reasoning, and reconstruction guided attention	Spatial	Around 150K videos across datasets	Improves accuracy in detecting face forgery, Handles unknown forgery patterns effectively, Analyzes discrepancies at different levels, Focuses on critical forgery traces for better classification, Outperforms existing methods on benchmark datasets
CORE [31]	CVPR 2022	FaceForensics++ [6]	Consistent Representation Learning, Consistency loss with different penalties, Xception network [30] for face forgery detection	Spatial	1000 real, 4000 fake videos	CORE (Consistent Representation Learning) framework is introduced for effective face forgery detection, Achieves state-of-the-art performance on the FF++ and Celeb-DF datasets. Demonstrates significant gains over existing methods, particularly in cross-dataset evaluations, Integrates flexibly with various backbone networks, showcasing its versatility
UCF [43]	CVPR 2023	FaceForensics++ [6]	Disentanglement framework to separate content and forgery features, Multi-task learning strategy to enhance detection performance, Primarily utilizes Xception architecture	Spatial	1,000 pristine videos	Superior performance across 27 detectors, reducing overfitting to specific artifacts, Separates content and fingerprint features using multi-task learning for enhanced detection, Achieves top AUC scores of 0.824 and 0.805 on CelebDF and DFDC datasets, Compatible with various architectures like Xception, ConvNeXt, and ResNet, Extensive ablation studies demonstrate improved generalization and resilience to biases

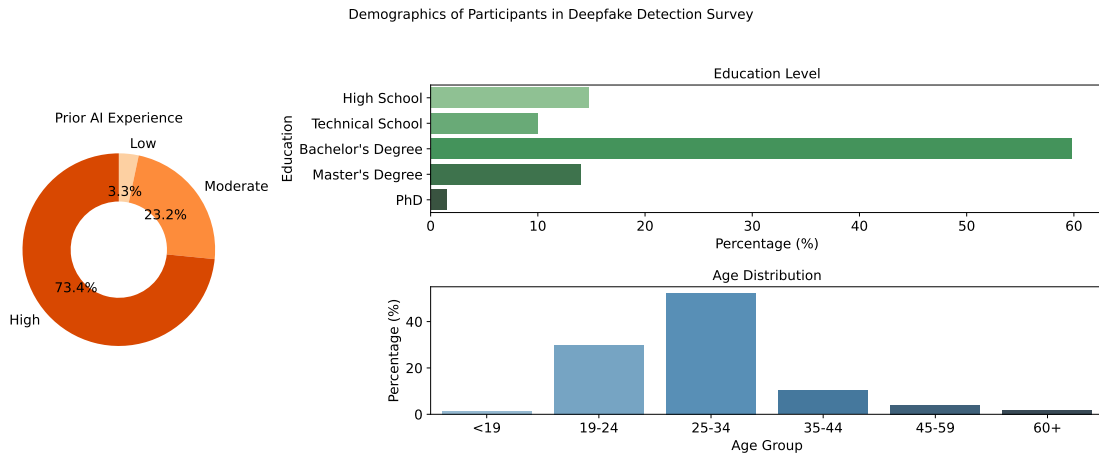


Figure 3.2: Demographic breakdown of participants in the deepfake detection survey. The majority of participants reported high prior AI experience (73.4%) and held at least a bachelor’s degree, with over half possessing a master’s degree. The largest age group was 25–34 years, indicating a relatively young and technologically proficient participant pool.

on the Qualtrics platform [48] consisting of three main sections: Demographic Information, Deepfake Detection – External Factors, and Deepfake Detection – Per Video. This questionnaire was administered to participants following their exposure to a set of real and AI-generated videos. The responses were collected to assess prior experience, subjective impressions, emotional responses, and decision-making criteria regarding the authenticity of video content.

Demographic Information

Participants were first asked to provide demographic information. They were requested to indicate their gender (Male, Female, Prefer Not to Say), race (American Indian or Native American, Asian / Pacific Islander / Asian Indian, Black or African American, Hispanic or Latino, White, or Other), and age group, categorized as: under 19, 19–24, 25–34, 35–44, 45–59, and 60 years or older. Educational background was also collected, with options ranging from below high school to graduate degree (e.g., M.S. or Ph.D.). Participants specified their country of residence by selecting from a list that included the United States, India, China, United Kingdom, other European countries, other Asian countries, Oceania countries, or none of the above. Lastly, to assess prior exposure to generative AI technologies, participants were asked about their experience with AI-generated content and algorithms, such as ChatGPT, GPT-4, DALL-E, Midjourney, and Stable Diffusion. Response options ranged from no prior knowledge to active use of both text- and image-based generative AI systems.

Deepfake Detection - External Factors

Q. No.	Question Text (5-point Likert Scale)
Q1	Sometimes I am amused by the products of deepfake technology.
Q2	People will learn to identify deepfakes progressively.
Q3	Deepfakes are a proof of technological advancements.
Q4	I am pretty capable of detecting deepfakes now.
Q5	I am always alert for detecting deepfake videos.
Q6	Most people need warnings, indications, or labels to identify a deepfake.
Q7	The majority of people are not alert to detect deepfake videos.

Deepfake Detection - Per Video (Example: anidiff_hr_001.mp4)

Q. No.	Question Text
Q1.1	How happy does the video make you? (5-point Likert Scale)
Q1.2	How sad does the video make you? (5-point Likert Scale)
Q1.3	How angry does the video make you? (5-point Likert Scale)
Q1.4	How frightened does the video make you? (5-point Likert Scale)
Q1.5	How impressive do you find this video? (5-point Likert Scale)
Q2	How likely will you share this video with others? (5-point Likert Scale)
Q3	Reasons to share this video (Select ALL that apply)
Q3 (Text)	Others, please specify: (Open text)
Q4	Reasons for NOT sharing this video (Select ALL that apply)
Q4 (Text)	Others, please specify: (Open text)
Q5	Is this video computer-generated or real? (Scale: Definitely Computer Generated (1.00) to Definitely Taken by Camera (0.00))
Q6	Reasons for your answer (Select ALL that apply)
Q6 (Text)	Others, please specify: (Open text)

A total of 271 participants successfully completed the survey: 148 identified as male, 121 as female, and 2 preferred not to say. Participants were recruited through Prolific [49] and were compensated for their participation. On average, participants spent approximately 20 minutes completing the survey. The demographic distribution of the participants is illustrated in Figure 3.2. The demographic data reveals a predominantly young, well-educated participant group, with 52.4% of respondents aged 25-34 and 59.8% holding a Bachelor's degree. This suggests that the survey attracted a tech-savvy audience, which may influence their ability to detect deepfakes. The inclusion of participants across

a range of ages and educational backgrounds provides a broader perspective on the human detection capabilities of deepfake videos. Participants were also categorized based on their prior experience with AI-generating algorithms, reflecting varying levels of familiarity with advanced AI tools. The participants were divided into three clusters: Low Experience, Moderate Experience, and High Experience. Cluster 1, Low Experience, consists of participants who have never heard of AI-generating algorithms like chatGPT, GPT-4, DALL-E, MidJourney, or StableDiffusion, accounting for 3.4% of the total participants. Cluster 2, Moderate Experience, includes those who have heard of chatGPT or GPT-4, but not DALL-E, MidJourney, or StableDiffusion, or have heard of at least one of these tools, comprising 23.2% of participants. Cluster 3, High Experience, consists of participants who have used chatGPT/GPT-4 or DALL-E/MidJourney/StableDiffusion, or both, representing 73.4% of the participants. This distribution highlights the prevalence of AI tool usage among participants, with a substantial majority having significant experience with these technologies.

The survey involved participants viewing 20 video clips, each lasting 5 seconds, with no accompanying audio. A total of 320 video stimuli were prepared for this survey, divided into eight categories. Seven of these categories consist of various types of deepfake videos, while the eighth category comprises real, unaltered videos. Each category includes 20 unique videos, and for each video, both low-resolution and high-resolution versions were presented to participants. Therefore, the total number of videos in the survey was calculated as follows:

$$\text{Total number of videos} = 8 \text{ categories} \times 20 \text{ videos} \times 2 \text{ resolutions} = 320.$$

This approach ensures that participants' ability to detect deepfakes is assessed across different resolutions and video types. These videos were part of an anonymous survey, which has been reviewed and approved by the Institutional Review Board (IRB) at the National University of Singapore (NUS). Informed consent was obtained from all participants prior to their participation in the study.

This methodological approach ensures a comprehensive evaluation of participants' ability to distinguish between real and deepfake videos across different types of manipulations and resolutions. The responses collected provide a robust dataset, which will facilitate a thorough analysis of human deepfake detection capabilities.

In this study, we evaluate the performance of deepfake detection models using a comprehensive set of metrics to capture various aspects of model behavior. The primary metrics include Area Under the Receiver Operating Characteristic Curve (AUC) and Average Precision (AP), which assess the overall ranking and classification capabilities of the models across thresholds. Additionally, we report Precision (P) and Recall (R) values to measure the models' ability to correctly identify true positives while minimizing false positives and false negatives. To further understand the sensitivity and decision-making tendencies of the detectors, we include Sensitivity Index (d') and Decision Criterion (C) metrics, which are derived from signal detection theory. This diverse suite of evaluation metrics allows for a detailed and multidimensional comparison of deepfake detection

models across various generation methods.

3.2.4 Multimodal Integration

This section outlines our unified framework for multimodal deepfake detection that integrates both visual and auditory modalities. We leverage state-of-the-art (SOTA) pre-trained models for each modality and investigate three distinct fusion strategies—Early Fusion, Late Fusion, and Cross-modal Attention—to explore their efficacy under realistic computational constraints.

Overview

Our proposed framework for multimodal deepfake detection is built around the strategic integration of video and audio information to exploit the complementary cues each modality offers. The system consists of the following core components, each chosen for specific technical advantages and suitability to the task. The overall pipeline is illustrated in Figure 3.3, which shows the flow from feature extraction to classification.

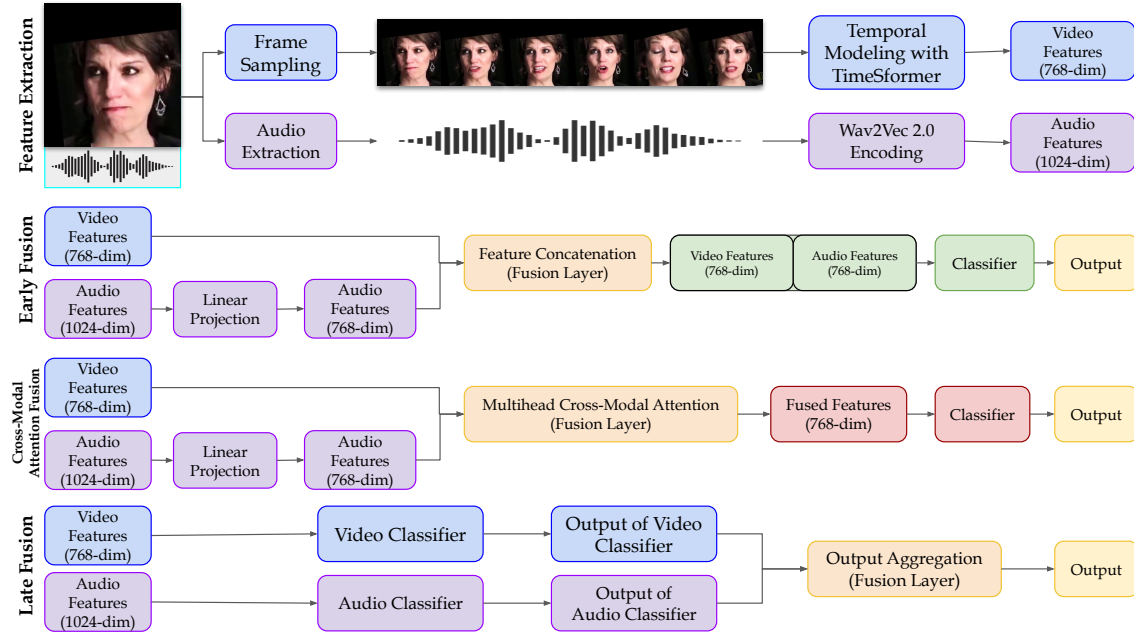


Figure 3.3: Overview of the multimodal deepfake detection framework. The pipeline begins with feature extraction, where video frames are sampled and audio is extracted from the input video. Temporal modeling is performed using TimeSformer for video and Wav2Vec 2.0 for audio to obtain modality-specific embeddings. Three fusion strategies are explored: (i) Early Fusion, where audio features are projected to match video dimensionality and concatenated before classification; (ii) Cross-Modal Attention Fusion, a form of early fusion enhanced with multi-head attention, where interactions between audio and visual features are learned and integrated into a fused representation before classification; and (iii) Late Fusion, where separate modality-specific classifiers are used and their outputs are aggregated.

Video Encoder: TimeSformer

We adopt the TimeSformer (Time-Space Transformer) [50], a transformer-based video model pretrained on the Kinetics-400 dataset [51]. This model is specifically designed for learning spatio-temporal representations from short video clips using attention mechanisms across both spatial and temporal dimensions.

Rationale for Choosing TimeSformer:

- **Efficiency over 3D CNNs:** Traditional 3D CNNs like I3D or C3D [52, 53, 54] are computationally intensive and less scalable. TimeSformer replaces convolutions with attention, reducing redundant computations.
- **Transfer learning strength:** Pretraining on Kinetics-400 provides a rich prior for motion and visual appearance, making it well-suited for tasks like deepfake detection, where artifacts may be subtle and temporally distributed.
- **Frame-based operation:** It works effectively with a limited number of frames (e.g., 8–16), allowing efficient training even on modest hardware.

Audio Encoder: Wav2Vec2.0

For audio, we utilize Wav2Vec2.0 (XLSR-53) [55], a self-supervised transformer model pretrained on 53 languages, enabling extraction of high-resolution, contextual audio representations from raw waveforms.

Rationale for Choosing Wav2Vec2.0:

- **Raw waveform input:** Avoids hand-crafted features (like MFCCs), reducing manual bias and capturing more nuanced distortions or synthesis artifacts in fake audio.
- **Temporal precision:** Its architecture is designed to preserve temporal information—essential for identifying mismatches in lip-sync or speech cadence.
- **Generalizability:** Pretraining on multilingual datasets increases robustness across various speech styles and audio qualities often found in manipulated videos.

Fusion Module: Strategy-Specific Design

We experiment with three fusion strategies—Early Fusion, Late Fusion, and Cross-modal Attention—each addressing a different hypothesis about how modalities interact:

- **Early Fusion** assumes shared representations are more informative when learned jointly.
- **Late Fusion** keeps modalities separate and combines outputs at decision level, which improves modularity and robustness.

- Cross-modal attention enables one modality (e.g., video) to act as a query and dynamically attend to the most relevant parts of another modality (e.g., audio) using a learned attention mechanism. Instead of selecting one modality over the other, this approach produces a fused representation that retains the semantic context of both modalities—modifying the video feature based on audio cues. This dynamic weighting allows the model to focus on inter-modal inconsistencies (such as mismatched lip movements and speech) that are often indicative of deepfakes, leading to more robust and context-aware detection.

Each fusion strategy reflects a trade-off between computational cost, model flexibility, and feature alignment capability. We compare them empirically to determine which approach is most effective under realistic conditions.

Classifier

A shared or modality-specific feed-forward binary classifier predicts the likelihood of a sample being real or fake. Its design is kept intentionally simple (2–3 layers with dropout) to minimize overfitting and highlight the contribution of the encoders and fusion strategy.

The architecture is trained and evaluated on a custom dataset comprising both real and fake videos, where each video has an accompanying audio track. Our goal is to detect manipulations in either or both modalities.

Dataset Collection

For the multimodal deepfake detection framework, we collected and utilized datasets that specifically focus on integrating visual and auditory modalities to enhance the detection of manipulated content. Our dataset consists of two primary configurations, which were selected to evaluate the impact of fake audio in the context of fake videos:

Fake Video with Fake Audio (11,000 samples): The dataset comprises manipulated videos paired with corresponding fake audio tracks. These pairs were sourced from the FakeAVCeleb [56] dataset.

The FakeAVCeleb [56] dataset was introduced to support multimodal deepfake detection, where both visual (face) and auditory (voice) signals can be real or synthetically manipulated. This dataset focuses on mimicking real-world deepfake threats where attackers may manipulate either or both modalities. Real videos were collected from YouTube interviews of 50 celebrities, amounting to around 500 high-quality videos. These videos featured frontal-facing shots and clean audio to support deepfake synthesis and alignment across modalities.

To generate visual deepfakes, models such as FaceSwap-GAN [20], DeepFaceLab [57], FSGAN [58], SimSwap [59], and FaceShifter [60] have been employed. These tools specialize in face-swapping techniques, allowing the replacement of one individual’s facial features with another while preserving the original video’s pose, expression, and lighting to maintain realism. In addition to face swapping, the dataset includes lip-sync deepfakes, where

the lips are animated to match an altered audio track, making it useful for evaluating deepfake detection models’ sensitivity to subtle inconsistencies.

For audio synthesis, two types of manipulations were used: Voice Conversion (VC) [61] and Text-to-Speech (TTS) [62]. The AutoVC [63] model was used to convert one speaker’s voice into another’s while preserving the content. Meanwhile, Tacotron2 [64] paired with MelGAN [65] was used for TTS-based fake audio generation, converting text into synthetic speech with high naturalness and speaker conditioning. As a result, FakeAVCeleb includes diverse fake combinations. This makes it especially valuable for studying cross-modal consistency and building robust multimodal deepfake detectors.

The fake video and fake audio pairs serve as the primary focus of our study, allowing us to investigate how well the integration of fake audio can enhance or hinder the detection of deepfake video content.

Real Video with Real Audio (11,000 samples): This baseline set consists of authentic video and audio pairs. These pairs were sourced from the following databases; 1M-Deepfakes Detection Challenge [66], and FakeAVCeleb [56]. It serves as the ground truth for comparison, providing a reference for classification and enabling the evaluation of the model’s ability to differentiate real content from manipulated videos.

The total dataset includes 22,000 samples, evenly distributed across these two categories. Each video is accompanied by its corresponding audio, ensuring that both modalities are analyzed simultaneously.

By using these varied data combinations, we aim to rigorously test the effectiveness of multimodal fusion strategies in the detection of deepfake content, specifically focusing on how the integration of fake audio can impact the accuracy of visual-based detection.

In this study, we primarily include two configurations: Fake Video with Fake Audio and Real Video with Real Audio. These two scenarios were selected to establish a clear distinction between completely manipulated and entirely authentic multimodal content. This binary setup allows us to analyze how the presence of synthetic audio paired with fake video affects detection performance. While the other two possible combinations (Fake Video with Real Audio and Real Video with Fake Audio) are also relevant and present more nuanced manipulation cases, they were excluded in this phase to maintain controlled conditions. We intend to incorporate these partial-manipulation cases in future research to further assess the system’s ability to handle asymmetric inconsistencies between modalities.

Dataset Preprocessing

Video Processing

Given the computational cost of full-frame video processing, we employ a clip-based strategy suitable for transformer-based video models such as TimeSformer. Each video is decoded using OpenCV and resized to 224×224 pixels. We extract frames at 25 FPS and fix the clip duration to 2.56 seconds, yielding exactly 64 frames per video. To ac-

commodate longer videos exceeding the 2.56-second constraint, we segment each video into multiple non-overlapping or partially overlapping clips of fixed duration (64 frames per clip). In cases where the remaining segment at the end of the video contains fewer than 64 frames, we apply zero-padding by repeating the last available frame to ensure uniform clip length. During inference, we aggregate predictions from all clips belonging to a single video using probability averaging, producing a robust video-level prediction while preserving the model’s confidence calibration. This clip-based approach enables efficient processing under computational constraints without compromising temporal consistency.

Each frame is transformed into a tensor, normalized to $[0, 1]$, and permuted to channel-first format (C, H, W) . The final video tensor is stacked and permuted to match the input shape expected by TimeSformer, resulting in a shape of $(T, C, H, W) = (16, 3, 224, 224)$ per video.

This deterministic and uniform frame sampling ensures reproducibility and enables training with modest hardware resources (e.g., 24–32 GB VRAM GPUs), striking a balance between temporal representation and computational efficiency.

Audio Processing

Audio is extracted directly from the video using FFmpeg, constrained to a duration of 2.56 seconds to align with the visual clip length. The extracted audio is converted to mono and resampled to 16 kHz, yielding a fixed-length waveform of 16,000 samples. For audio samples shorter than 2.56 seconds, we apply zero-padding with silence (zeros) to achieve a fixed length of 16,000 samples. During inference, audio segments are chunked into non-overlapping 2.56-second clips to maintain alignment with the corresponding visual segments. If multiple audio clips are derived from a single video, their prediction scores are averaged to produce a consistent and reliable audio-level decision. This ensures temporal synchronization with the video stream and enables efficient multimodal fusion during both training and evaluation.

The processed waveform is passed to Wav2Vec2.0 XLSR-53, a multilingual transformer-based model pre-trained on 53 languages. We utilize its intermediate hidden representations rather than just the final logits, allowing for richer feature extraction and flexible fusion strategies with visual modalities.

Fusion Strategies

We explore three fusion strategies, each reflecting a different philosophy of multimodal integration. All strategies use the same pretrained backbones for fair comparison.

Early Fusion

To enable the model to learn joint representations that capture cross-modal correlations, we combine the features from each modality prior to classification. This early fusion approach facilitates deeper integration of multimodal information, allowing the network

to model complex interdependencies and complementary signals across modalities during the learning process.

Design:

- Video features (768-dim) and audio features (1024-dim) are extracted.
- Audio features are projected via a fully connected layer to match video dimensions (768-dim).
- The two vectors are concatenated and passed through a shared classifier.

Early fusion assumes that deepfake cues in video and audio are mutually reinforcing and can be more effectively modeled in a shared embedding space. However, it increases the risk of feature misalignment and may suffer from overfitting on small datasets.

Late Fusion

To reduce modality-specific interference and preserve the unique characteristics of each modality, we treat each modality independently during processing. Rather than fusing features early, we combine the decisions or outputs from each modality at a later stage. This decision-level fusion strategy promotes modularity and allows each modality to contribute its strengths to the final prediction without being distorted by incompatible representations.

Design:

- Separate classifiers are trained on video and audio representations.
- Each branch outputs a confidence score, and final prediction is the mean of the two scores.

Late fusion avoids direct feature alignment issues and allows each modality to specialize. It's more robust when one modality is corrupted or less informative but may underutilize inter-modal correlations.

Cross-modal Attention Fusion

This module is designed to mimic human cognitive strategies where cues from one sensory channel—like speech—can guide our visual attention and interpretation. In the context of deepfake detection, such cross-modal reasoning is crucial because many deepfakes exploit misalignments between audio and video that are difficult to detect when modalities are treated in isolation.

As illustrated in Figure 3.4, the model takes visual features extracted from TimeS-former and audio features from Wav2Vec 2.0. Since the two modalities operate in different representational spaces, we first project the audio features to match the dimensionality of

video features. Then, a multi-head attention mechanism is applied where video features serve as Queries, and audio features act as both Keys and Values.

This formulation enables the video modality to selectively attend to semantically aligned or temporally relevant segments in the audio stream. In other words, the model learns to enhance its visual representation by focusing on those parts of the audio that provide meaningful context—such as speech onsets, pauses, emotional tone, or unnatural prosody. For example, if the lip movement doesn’t match the spoken syllables or if the emotional tone of the voice does not align with the facial expression, the attention mechanism helps surface these cross-modal discrepancies.

The result is an audio-guided visual representation that is more robust and context-aware. This enriched representation is then passed through a classifier to determine the authenticity of the content. By allowing one modality to guide and refine another, the model becomes significantly better at detecting subtle manipulation artifacts, particularly those that exploit the temporal and semantic disconnect between what is seen and what is heard.

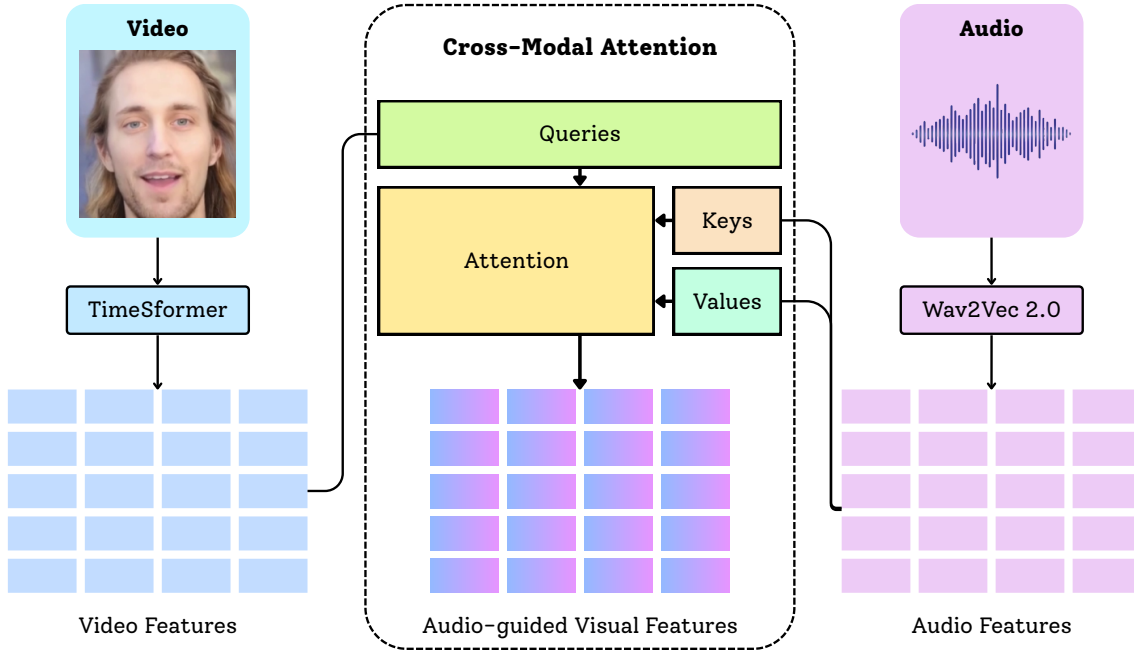


Figure 3.4: Illustration of the proposed Cross-Modal Attention mechanism for multi-modal deepfake detection. Visual features extracted from TimeSformer act as Queries, while audio features from Wav2Vec 2.0 serve as both Keys and Values. This setup allows the video modality to dynamically attend to the most informative segments of the audio stream. As a result, audio cues guide the refinement of visual representations, enabling the model to highlight inconsistencies—such as lip-sync mismatches or emotion desynchronization—that are commonly found in deepfakes.

The proposed multimodal deepfake detection framework leverages both audio and visual data to enhance detection accuracy. By integrating video-based temporal modeling using TimeSformer and audio feature extraction with Wav2Vec 2.0, the framework accom-

modates diverse fusion strategies: Early Fusion, Late Fusion, and Cross-modal Attention. This approach not only addresses the challenges posed by unimodal detection methods but also provides a more resilient solution to evolving deepfake techniques. Future work could explore optimizing the fusion mechanisms and incorporating other modalities to further strengthen the detection robustness.

3.3 Summary

This chapter outlined the essential requirements for evaluating deepfake detection models against modern synthesis methods. By establishing a structured methodology that incorporates both algorithmic and human evaluations, this study aims to bridge the gap between generation and detection capabilities. The outlined requirements highlight the importance of diverse data, robust detection models, human-centric analysis, and multimodal integration, setting the foundation for a comprehensive empirical assessment in subsequent chapters.

Chapter 4

Implementation, Evaluation, and Results

This chapter presents the implementation details of our deepfake detection system and the results obtained from evaluating various models. It includes the environment setup, the process of generating synthetic videos, deepfake detection methodologies, and the comparative performance analysis of unimodal and multimodal approaches.

4.1 Experimental Environment and Setup

The experiments were conducted on a Linux-based system, utilizing a single NVIDIA GeForce RTX 3090 GPU with 24 GB of memory. The system is equipped with an AMD EPYC 7313P 16-core processor, providing robust computational power. The environment for the experiments was set up using Anaconda, with the specific environment named deepFake, which includes various essential packages such as PyTorch, OpenCV, and TensorFlow. The system has 256 GB of RAM and a 32 GB swap space, ensuring efficient handling of large datasets. The CUDA version 12.3 and the appropriate NVIDIA drivers were installed to enable GPU acceleration, facilitating faster model training and evaluation.

4.2 Hyperparameter Settings

The proposed multimodal deepfake detection framework was trained using a consistent set of hyperparameters designed to ensure stable convergence and robust performance across different fusion strategies. The Adam optimizer was utilized due to its adaptive learning rate capabilities and proven effectiveness in training deep neural networks. A fixed learning rate of $1e-4$ was applied throughout the training process, offering a balance between convergence speed and training stability. The models were trained with a batch size of 8, which was selected considering GPU memory limitations associated with simultaneously processing high-resolution video frames and long audio sequences. Training was conducted

for 30 epochs, as this duration was empirically found to be sufficient for model convergence without overfitting. Binary Cross-Entropy (BCE) loss was adopted as the objective function, aligning with the binary nature of the classification task (real vs. fake). A dropout rate of 0.3 was introduced within the classification head to mitigate overfitting and enhance generalization. For early fusion, audio and video features were projected into a shared embedding space of 768 dimensions before concatenation. In the cross-modal attention module, a multi-head attention mechanism was employed with 8 attention heads and an embedding size of 768. In this configuration, video features served as the query input, while audio features were used as keys and values—enabling the model to guide visual attention based on relevant audio cues. Input data was also standardized: video frames were resized to 224×224 pixels, and audio segments were temporally aligned with the video clips, preserving their native sampling rate. These hyperparameter choices were kept consistent across all experiments to ensure fair comparison among the early fusion, late fusion, and cross-modal attention strategies.

4.3 Performance Evaluation

4.3.1 Synthetic Video Generation and Dataset Preparation

We used open-source models from the original authors to generate videos with source images and driving videos from the VoxCeleb dataset, ensuring transparency, reproducibility, and use of state-of-the-art techniques. Selection was based on video quality, model/code availability, and publication date—prioritizing high-quality synthesis for accurate evaluation, consistent implementation, and relevance to recent advancements. This curated set supported robust comparisons across both algorithmic and human detection.

4.3.2 Unimodal Deepfake Detection Analysis

This section presents a comparative analysis of deepfake detection performance between automated algorithms and human participants. For model evaluation, we employed a wide range of metrics: Area Under the ROC Curve (AUC) [67] and Average Precision (AP) to assess overall classification performance; Precision and Recall to evaluate the accuracy in identifying true positives and negatives; and d' and C from signal detection theory [68] to analyze discriminability and response bias. This robust framework supports a detailed comparison of both human and algorithmic detection capabilities across various synthesis methods and quality levels.

Since humans do not provide probabilistic outputs, participants rated each video’s authenticity on a five-point scale ranging from “Definitely computer generated” to “Definitely taken by camera.” These responses were converted to probability scores, with values from 1.00 to 0.00, and averaged per video to enable comparison with model predictions.

Results are summarized in Tables 4.1, 4.2, and 4.3, covering AUC, AP, Precision, Recall, and signal detection theory metrics such as d' -prime and criterion, along with the

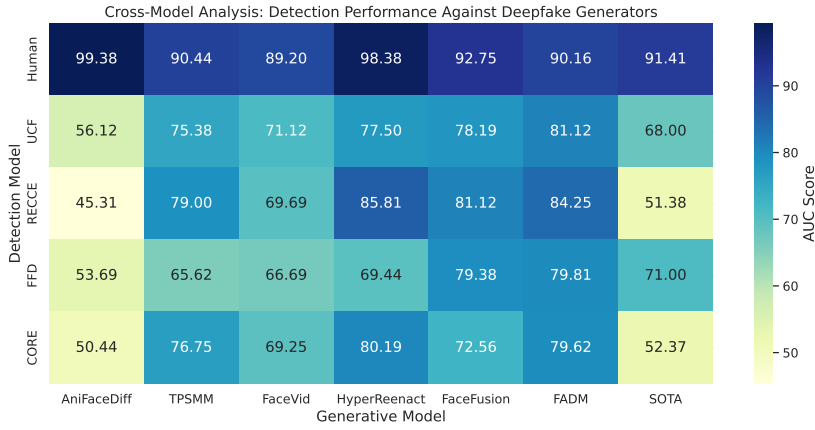


Figure 4.1: Cross-model detection performance (AUC scores) of the four best-performing automated deepfake detection algorithms compared to human participants, evaluated across generative models. Humans consistently outperform all automated methods across all generators, particularly excelling against AniFaceDiff [11] and HyperReenact [14]. Humans demonstrate a strong capability to detect deepfakes generated by state-of-the-art (SOTA) models such as SyncTalk [4], DreamTalk [15], Diffused Heads [16], and VASA-1 [3], highlighting their perceptual advantage over current automated detection systems. The results highlight the current limitations of detection algorithms in generalizing across diverse deepfake generation techniques.

mean and standard deviation across models and manipulation types.

Human evaluators outperformed automated methods, achieving a mean AUC of 93.10 and a mean AP of 94.81, with low standard deviations that reflect high accuracy and consistency. Among models, FFD [42] and UCF [43] showed relatively strong performance. For instance, FFD reached an AUC of 69.38 and an AP of 71.00. In contrast, MesoNet [5] demonstrated much higher variability, signaling inconsistent detection performance. These results are summarized in Table 4.1 and illustrated in Figure 4.1.

Table 4.2 further highlights the robustness of human detection. Human evaluators showed a high mean precision of 98.93 and a moderate recall of 70.72. While recall was more variable, the consistently high precision suggests that humans tend to be cautious yet accurate. Capsule [41] and RECCE [18] achieved competitive precision but suffered from elevated variability in recall. MesoNet’s uniformly perfect precision and recall scores of 100 percent appear anomalous and may indicate artifacts or thresholding issues that warrant further examination.

Table 4.3 frames detection performance using signal detection theory, providing complementary insights beyond AUC and AP. While AUC and AP measure overall classification performance, SDT metrics such as d' and C capture the observer’s sensitivity (discriminability) and decision bias, offering a more nuanced understanding of both human and model behavior in deepfake detection tasks. Human participants achieved a high mean d' -prime of 7.01 with low variability, reflecting strong and consistent discrimination

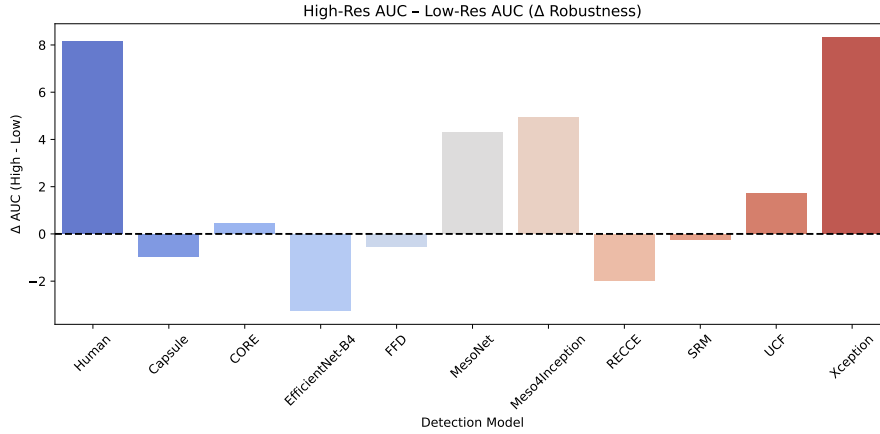


Figure 4.2: ΔAUC (High-Res – Low-Res) for each detection model, representing robustness to video resolution changes. Each AUC value is the average performance across multiple generative models (e.g., different deepfake sources). Positive Δ values indicate better performance on high-resolution videos. Xception [6] and Human detectors show the largest gains (>8 AUC points), while EfficientNet-B4 [17] and RECCE [18] perform worse on high-res inputs. The overall average ΔAUC across all models is $+1.90$, suggesting that high-resolution videos modestly improve detection performance on average.

between real and fake videos. Their average criterion of 2.86 suggests a cautious response bias, aligning with their high precision. Automated models showed much weaker discrimination, with CORE [31] and EfficientNet-B4 [17] scoring d-primes below 1 and greater variability overall. MesoNet [5] again displayed unreliable results, with a d-prime of 0 and an implausibly extreme criterion of -6.36.

To evaluate robustness to resolution, we analyzed detection performance on high- versus low-resolution videos. The Tables 4.4, 4.5 provide the complete quantitative results corresponding to the visual summaries presented Figure 4.2. These tables report the performance of various deepfake detection models evaluated against multiple generation methods, with metrics including Area Under the Curve (AUC) and Average Precision (AP). All evaluations were conducted on 5-second video clips, with results stratified by resolution to assess robustness across different video qualities. Table 4.4 presents the results on high-resolution videos, while Table 4.5 reports performance on low-resolution videos. These detailed metrics support the main analysis and enable deeper examination of how detection performance varies across models, generation methods, and resolution levels. As shown in Figure 4.2, we computed the change in AUC (ΔAUC) for each model. Human and Xception [6] detectors showed the largest improvements, both gaining over 8 AUC points on high-resolution inputs. Meanwhile, EfficientNet-B4 and RECCE [18] performed worse on high-res data, hinting at overfitting or limited generalization. On average, detection models showed a modest improvement in 1.90 AUC points, emphasizing that video quality can significantly influence the effectiveness of deepfake detection.

Table 4.1: Comparison of Detection and Generation Performance (AUC and AP values) on 5-Second Videos for Various Deepfake Detection Models (ROWS) and Generators (COLUMNS). The bolded values highlight the best overall detection model across rows and the most challenging deepfake generator across columns.

Model	AniFaceDiff [11]		TPSMM [2]		FaceVid [1]		HyperReenact [14]		FaceFusion [12]		FADM [13]		SOTA [3, 4, 15, 16]		Detection Performance: Mean $\pm$ Std	
	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP
Human	99.38	99.43	90.44	93.85	89.20	91.71	98.38	98.92	92.75	94.65	90.16	92.67	91.41	92.47	93.10 $\pm$ 4.11	94.81 $\pm$ 3.13
Capsule [41]	49.94	49.70	71.31	72.74	78.56	77.09	69.19	65.98	65.56	61.67	83.94	81.6	47.38	51.66	66.55 $\pm$ 13.67	65.78 $\pm$ 12.26
CORE [31]	50.44	53.86	76.75	75.43	69.25	70.51	80.19	81.38	72.56	74.56	79.62	79.05	52.37	60.60	68.74 $\pm$ 12.46	70.77 $\pm$ 10.06
EfficientNet-B4 [17]	50.19	55.03	63.06	69.97	71.06	74.23	68.56	70.64	67.44	72.84	78.00	79.92	51.37	66.00	64.24 $\pm$ 10.24	69.80 $\pm$ 7.80
FFD [42]	53.69	54.57	65.62	67.21	66.69	71.34	69.44	71.47	79.38	80.17	79.81	81.66	71.00	70.58	69.38 $\pm$ 8.93	71.00 $\pm$ 8.97
MesoNet [5]	40.88	48.44	32.88	39.07	42.19	45.43	83.00	84.43	47.06	52.87	39.50	48.00	62.62	66.85	49.73 $\pm$ 17.34	55.01 $\pm$ 15.54
Meso4Inception [5]	53.62	53.15	60.88	58.77	69.12	69.08	57.00	54.82	63.62	61.16	73.88	67.71	55.75	56.71	61.98 $\pm$ 7.42	60.20 $\pm$ 6.18
RECCE [18]	45.31	52.33	79.00	82.95	69.69	74.86	85.81	88.42	81.12	85.93	84.25	86.83	51.38	63.57	70.94 $\pm$ 16.37	76.41 $\pm$ 13.74
SRM [7]	43.50	47.36	66.75	64.64	63.87	62.69	70.19	65.75	73.81	75.49	76.88	73.08	49.00	58.72	63.43 $\pm$ 12.59	63.96 $\pm$ 9.35
UCF [43]	56.12	57.49	75.38	75.83	71.12	74.38	77.50	75.71	78.19	78.89	81.12	77.43	68.00	71.67	72.49 $\pm$ 8.47	73.06 $\pm$ 7.23
Xception [6]	43.62	45.07	66.12	66.58	69.31	65.30	74.06	67.68	76.25	78.12	81.12	77.54	68.50	66.20	68.43 $\pm$ 12.08	66.64 $\pm$ 10.95
Generation Performance: Mean $\pm$ Std	53.34 $\pm$ 16.03	56.04 $\pm$ 14.86	68.02 $\pm$ 14.44	69.73 $\pm$ 13.92	69.10 $\pm$ 11.22	70.60 $\pm$ 11.23	75.76 $\pm$ 10.99	75.02 $\pm$ 12.42	72.52 $\pm$ 11.75	74.21 $\pm$ 11.89	77.12 $\pm$ 13.19	76.86 $\pm$ 11.58	60.80 $\pm$ 13.23	65.91 $\pm$ 10.65		

Table 4.2: Comparison of Detection and Generation Performance (P, Precision and R, Recall Values) on 5-Second Videos for Various Deepfake Detection Models (ROWS) and Generators (COLUMNS). The bolded values highlight the best overall detection model across rows and the most challenging deepfake generator across columns.

Model	AniFaceDiff [11]		TPSMM [2]		FaceVid [1]		HyperReenact [14]		FaceFusion [12]		FADM [13]		SOTA [3, 4, 15, 16]		Detection Performance: Mean $\pm$ Std	
	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R
Human	92.50	96.10	100.00	67.50	100.00	58.97	100.00	95.00	100.00	75.00	100.00	55.00	100.00	47.50	98.93 $\pm$ 2.83	70.72 $\pm$ 19.09
Capsule [41]	50.00	47.50	61.22	75.00	64.81	87.50	60.42	72.50	59.57	70.00	66.07	92.50	42.42	35.00	57.79 $\pm$ 8.53	68.57 $\pm$ 20.66
CORE [31]	43.75	17.50	67.86	47.50	66.67	45.00	70.97	55.00	70.00	52.50	73.53	62.50	60.87	35.00	64.81 $\pm$ 10.11	45.00 $\pm$ 14.86
EfficientNet-B4 [17]	52.63	25.00	65.38	42.50	67.86	47.50	64.00	40.00	66.67	45.00	70.00	52.50	60.87	35.00	63.92 $\pm$ 5.76	41.07 $\pm$ 9.00
FFD [42]	52.63	25.00	64.00	40.00	66.67	45.00	66.67	45.00	71.88	57.50	74.29	65.00	66.67	45.00	66.12 $\pm$ 6.92	46.07 $\pm$ 12.74
MesoNet [5]	50.00	100.00	50.00	100.00	50.00	100.00	50.00	100.00	50.00	100.00	50.00	100.00	50.00	100.00	50.00 $\pm$ 0.00	100.00 $\pm$ 0.00
Meso4Inception [5]	50.91	70.00	55.74	85.00	57.14	90.00	53.45	77.50	55.00	82.50	59.09	97.50	54.24	80.00	55.08 $\pm$ 2.63	83.21 $\pm$ 8.86
RECCE [18]	62.50	12.50	86.96	50.00	81.25	32.50	88.00	55.00	88.46	57.50	88.00	55.00	76.92	25.00	81.73 $\pm$ 9.53	41.07 $\pm$ 17.73
SRM [7]	46.43	32.50	63.41	65.00	59.46	55.00	65.91	72.50	66.67	75.00	67.39	77.50	48.28	35.00	59.65 $\pm$ 8.82	58.93 $\pm$ 18.76
UCF [43]	50.00	12.50	73.68	35.00	73.68	35.00	66.67	25.00	75.00	37.50	76.19	40.00	66.67	25.00	68.84 $\pm$ 9.16	30.00 $\pm$ 9.68
Xception [6]	38.89	17.50	68.57	60.00	69.44	62.50	66.67	55.00	72.50	72.50	73.81	77.50	66.67	55.00	65.22 $\pm$ 11.93	57.14 $\pm$ 19.44
Generation Performance: Mean $\pm$ Std	49.77 $\pm$ 6.20	36.00 $\pm$ 28.73	65.68 $\pm$ 10.01	60.00 $\pm$ 21.25	65.70 $\pm$ 8.73	60.00 $\pm$ 24.21	65.28 $\pm$ 10.29	59.75 $\pm$ 21.46	67.58 $\pm$ 10.92	65.00 $\pm$ 18.67	69.84 $\pm$ 10.27	72.00 $\pm$ 20.41	59.36 $\pm$ 10.52	47.00 $\pm$ 24.74		

Table 4.3: Comparison of Detection and Generation Performance (d' , Sensitivity Index and C, Decision Criterion Values) on 5-Second Videos for Various Deepfake Detection Models (ROWS) and Generators (COLUMNS). The bolded values highlight the best overall detection model across rows and the most challenging deepfake generator across columns.

Model	AniFaceDiff [11]		TPSMM [2]		FaceVid [1]		HyperReenact [14]		FaceFusion [12]		FADM [13]		SOTA [3, 4, 15, 16]		Detection Performance: Mean $\pm$ Std	
	d'	C	d'	C	d'	C	d'	C	d'	C	d'	C	d'	C	d'	C
Human	7.80	2.46	6.82	2.95	6.59	3.07	8.01	2.36	7.04	2.84	6.49	3.12	6.30	3.21	7.01 $\pm$ 0.66	2.86 $\pm$ 0.33
Capsule [41]	0.00	0.06	0.74	-0.31	1.21	-0.54	0.66	-0.27	0.59	-0.23	1.50	-0.69	-0.32	0.22	0.63 $\pm$ 0.63	-0.25 $\pm$ 0.32
CORE [31]	-0.18	0.85	0.69	0.41	0.63	0.44	0.88	0.31	0.82	0.35	1.07	0.22	0.37	0.57	0.61 $\pm$ 0.41	0.45 $\pm$ 0.21
EfficientNet-B4 [17]	0.08	0.71	0.57	0.47	0.69	0.41	0.50	0.50	0.63	0.44	0.82	0.35	0.37	0.57	0.52 $\pm$ 0.24	0.49 $\pm$ 0.12
FFD [42]	0.08	0.71	0.50	0.50	0.63	0.44	0.63	0.44	0.94	0.28	1.14	0.19	0.63	0.44	0.65 $\pm$ 0.33	0.43 $\pm$ 0.17
MesoNet [5]	0.00	-6.36	0.00	-6.36	0.00	-6.36	0.00	-6.36	0.00	-6.36	0.00	-6.36	0.00	-6.36	0.00 $\pm$ 0.00	-6.36 $\pm$ 0.00
Meso4Inception [5]	0.07	-0.49	0.58	-0.75	0.83	-0.87	0.30	-0.60	0.48	-0.69	1.51	-1.21	0.39	-0.65	0.59 $\pm$ 0.47	-0.75 $\pm$ 0.23
RECCE [18]	0.29	1.29	1.44	0.72	0.99	0.95	1.57	0.66	1.63	0.63	1.57	0.66	0.77	1.06	1.18 $\pm$ 0.51	0.85 $\pm$ 0.25
SRM [7]	-0.14	0.39	0.70	-0.03	0.44	0.10	0.92	-0.14	0.99	-0.18	1.07	-0.22	-0.07	0.35	0.56 $\pm$ 0.5	0.04 $\pm$ 0.25
UCF [43]	0.00	1.15	0.77	0.77	0.77	0.77	0.48	0.91	0.83	0.73	0.90	0.70	0.48	0.91	0.60 $\pm$ 0.31	0.85 $\pm$ 0.16
Xception [6]	-0.34	0.77	0.85	0.17	0.92	0.14	0.72	0.24	1.20	0.00	1.35	-0.08	0.72	0.24	0.77 $\pm$ 0.55	0.21 $\pm$ 0.27
Generation Performance: Mean $\pm$ Std	-0.01 $\pm$ 0.17	-0.09 $\pm$ 2.26	0.68 $\pm$ 0.36	-0.44 $\pm$ 2.13	0.71 $\pm$ 0.33	-0.45 $\pm$ 2.15	0.67 $\pm$ 0.42	-0.43 $\pm$ 2.13	0.81 $\pm$ 0.44	-0.50 $\pm$ 2.10	1.09 $\pm$ 0.46	-0.64 $\pm$ 2.09	0.33 $\pm$ 0.36	-0.27 $\pm$ 2.19		

Table 4.4: Comparison of Detection and Generation Performance (AUC and AP values) on 5-second High Resolution Videos for Various Deepfake Detection Models (ROWS) and Generators (COLUMNS). The bolded values highlight the best overall detection model across rows and the most challenging deepfake generator across columns.

Model	AniFaceDiff [11]		TPSMM [2]		FaceVid [1]		HyperReenact [14]		FaceFusion [12]		FADM [13]		SOTA [3, 4, 15, 16]		Detection Performance: Mean $\pm$ Std	
	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP
Human	99.75	99.76	93.50	95.67	91.58	93.93	98.00	98.45	95.75	96.63	99.25	99.25	97.00	96.87	96.40 $\pm$ 3.01	97.22 $\pm$ 2.07
Capsule [41]	47.5	48.59	74.00	75.67	80.00	78.19	65.50	62.47	66.00	65.91	85.75	81.38	57.75	60.24	68.07 $\pm$ 13.11	67.49 $\pm$ 11.63
CORE [31]	46.25	49.33	84.50	82.03	70.50	67.00	84.25	81.22	69.50	71.72	83.00	79.72	65.50	70.02	71.93 $\pm$ 13.79	71.58 $\pm$ 11.45
EfficientNet-B4 [17]	45.50	46.88	65.50	71.09	69.25	70.66	69.00	66.52	67.00	73.85	77.25	74.41	60.00	70.82	64.79 $\pm$ 9.94	67.75 $\pm$ 9.55
FFD [42]	50.25	50.40	70.00	70.66	63.75	66.59	67.75	66.24	81.50	82.36	79.00	81.48	76.50	74.31	69.82 $\pm$ 10.72	70.28 $\pm$ 10.88
MesoNet [5]	41.50	49.05	35.25	41.04	46.00	50.59	85.50	86.78	46.00	51.20	41.25	50.85	67.75	73.62	51.89 $\pm$ 18.02	57.59 $\pm$ 16.28
Meso4Inception [5]	53.50	54.54	67.75	64.51	72.00	71.50	57.50	53.54	67.25	64.56	80.75	74.19	71.75	68.2	67.21 $\pm$ 9.21	64.43 $\pm$ 7.91
RECCE [18]	39.75	45.38	81.75	84.56	68.00	71.00	83.75	84.47	77.00	84.98	85.50	86.81	62.50	70.34	71.18 $\pm$ 16.24	75.36 $\pm$ 14.91
SRM [7]	36.75	43.62	73.00	68.57	61.50	59.62	67.50	61.24	74.00	78.12	77.50	73.73	57.00	62.59	63.89 $\pm$ 13.98	63.93 $\pm$ 11.25
UCF [43]	53.50	53.77	81.50	82.36	68.750	72.70	76.75	74.08	79.00	81.91	83.00	80.33	75.50	77.30	74.00 $\pm$ 10.16	74.64 $\pm$ 9.93
Xception [6]	41.00	44.10	75.00	74.09	71.25	65.01	82.00	74.16	77.50	81.34	85.50	78.83	78.00	72.30	72.89 $\pm$ 14.8	69.98 $\pm$ 12.53
Generation Performance: Mean $\pm$ Std	50.48 $\pm$ 17.24	53.22 $\pm$ 15.83	72.89 $\pm$ 14.91	73.66 $\pm$ 13.94	69.33 $\pm$ 11.22	69.71 $\pm$ 10.86	76.14 $\pm$ 11.75	73.56 $\pm$ 13.23	72.77 $\pm$ 12.34	75.68 $\pm$ 12.18	79.80 $\pm$ 14.14	78.27 $\pm$ 11.59	69.93 $\pm$ 11.70	72.42 $\pm$ 9.50		

Table 4.5: Comparison of Detection and Generation Performance (AUC and AP values) on 5-second Low Resolution Videos for Various Deepfake Detection Models (ROWS) and Generators (COLUMNS). The bolded values highlight the best overall detection model across rows and the most challenging deepfake generator across columns.

Model	AniFaceDiff [11]		TPSMM [2]		FaceVid [1]		HyperReenact [14]		FaceFusion [12]		FADM [13]		SOTA [3, 4, 15, 16]		Detection Performance: Mean $\pm$ Std	
	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP
Human	98.50	98.75	87.50	92.00	84.00	87.80	99.75	99.76	88.50	92.09	76.00	83.14	83.50	87.10	88.25 $\pm$ 7.83	91.52 $\pm$ 5.66
Capsule [41]	53.25	53.83	70.75	73.91	84.25	82.80	81.00	74.77	67.50	62.72	89.50	87.51	37.00	45.87	69.04 $\pm$ 18.59	68.77 $\pm$ 15.25
CORE [31]	55.75	60.86	73.25	75.57	79.00	79.86	87.00	88.94	78.50	79.81	87.75	86.94	39.25	54.02	71.50 $\pm$ 17.80	75.14 $\pm$ 13.06
EfficientNet-B4 [17]	58.75	63.26	63.50	72.50	78.75	82.00	75.75	78.22	69.25	75.25	87.50	89.45	42.75	62.52	68.04 $\pm$ 14.72	74.74 $\pm$ 9.73
FFD [42]	58.75	61.58	62.25	65.99	72.00	77.33	75.25	78.70	77.25	78.77	81.50	82.79	65.50	67.42	70.36 $\pm$ 8.39	73.23 $\pm$ 8.07
MesoNet [5]	41.50	50.08	29.25	38.74	39.25	44.48	81.50	83.64	47.75	55.77	36.50	46.88	57.50	61.65	47.61 $\pm$ 17.39	54.46 $\pm$ 14.88
MesoInception [5]	54.25	55.22	57.50	60.17	75.25	73.81	66.25	62.66	65.75	63.69	77.25	69.69	39.75	49.38	62.29 $\pm$ 13.01	62.09 $\pm$ 8.27
RECCE [18]	52.00	61.43	76.25	83.68	77.75	82.03	95.75	96.32	83.25	86.28	86.75	89.23	40.25	58.94	73.14 $\pm$ 19.83	79.70 $\pm$ 14.12
SRM [7]	49.75	52.58	61.25	63.63	69.00	71.15	77.50	75.97	72.50	75.59	77.75	78.33	41.00	56.35	64.11 $\pm$ 14.19	67.66 $\pm$ 10.24
UCF [43]	60.00	62.64	68.50	69.71	75.75	78.43	82.25	79.32	77.50	75.45	81.50	76.93	60.50	67.10	72.29 $\pm$ 9.38	72.80 $\pm$ 6.37
Xception [6]	45.75	49.50	56.00	61.61	70.50	72.57	69.75	68.14	72.75	76.01	78.25	78.64	59.00	62.29	64.57 $\pm$ 11.36	66.97 $\pm$ 10.05
Average - Generation Model	57.11 $\pm$ 14.85	60.88 $\pm$ 13.57	64.18 $\pm$ 14.76	68.86 $\pm$ 13.82	73.23 $\pm$ 12.30	75.66 $\pm$ 11.46	81.07 $\pm$ 10.12	80.59 $\pm$ 11.12	72.77 $\pm$ 10.71	74.68 $\pm$ 10.51	78.20 $\pm$ 14.65	79.05 $\pm$ 12.26	51.45 $\pm$ 14.85	61.15 $\pm$ 10.97		

Table 4.6: Comparison of Detection and Generation Performance (AUC and AP values) on Full Length Videos for Various Deepfake Detection Models (ROWS) and Generators (COLUMNS). The bolded values highlight the best overall detection model across rows and the most challenging deepfake generator across columns.

Model	AniFaceDiff [11]		TPSMM [2]		FaceVid [1]		HyperReenact [14]		FaceFusion [12]		FADM [13]		SOTA [3, 4, 15, 16]		Detection Performance: Mean $\pm$ Std	
	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP
Capsule [41]	79.06	81.83	92.97	93.75	69.61	72.84	59.00	61.56	62.74	38.56	82.40	88.16	43.00	57.51	69.83 $\pm$ 16.66	70.6 $\pm$ 19.39
CORE [31]	87.13	90.39	99.12	99.17	69.50	69.34	85.22	86.87	62.43	41.98	80.02	84.74	59.50	67.30	77.56 $\pm$ 14.38	77.11 $\pm$ 19.18
EfficientNet-B4 [17]	85.96	88.34	91.37	92.96	64.12	66.80	39.46	47.52	43.81	34.68	70.43	79.71	48.25	67.56	63.34 $\pm$ 20.52	68.22 $\pm$ 21.23
FFD [42]	82.60	86.78	94.78	96.04	60.73	61.67	76.99	81.15	48.52	32.2	64.98	72.28	68.25	72.94	70.98 $\pm$ 15.22	71.87 $\pm$ 20.07
MesoNet [5]	51.05	52.02	44.13	45.53	37.32	43.75	58.35	54.90	39.98	24.17	42.08	59.04	61.00	60.91	47.70 $\pm$ 9.25	48.62 $\pm$ 12.53
MesoInception [5]	54.94	57.87	72.59	68.66	62.60	59.88	30.88	40.98	24.58	19.68	50.16	60.12	23.75	38.98	45.64 $\pm$ 19.41	49.45 $\pm$ 16.98
RECCE [18]	88.72	91.81	96.22	97.01	63.93	68.08	77.90	83.47	64.87	51.05	72.96	82.01	51.75	63.55	73.76 $\pm$ 15.31	76.71 $\pm$ 16.43
SRM [7]	85.01	89.02	87.52	89.55	56.55	57.03	81.22	84.69	64.05	49.10	67.72	77.40	47.50	59.43	69.94 $\pm$ 15.20	72.32 $\pm$ 16.80
UCF [43]	87.32	88.95	98.41	98.53	64.74	68.85	73.38	77.18	58.37	43.81	79.42	59.50	69.79	73.49 $\pm$ 14.8	75.22 $\pm$ 17.37	
Xception [6]	87.74	90.10	98.61	98.87	69.99	69.85	87.24	90.3	74.10	60.93	81.50	86.07	83.50	85.31	83.24 $\pm$ 9.44	83.06 $\pm$ 13.10
Generation Performance: Mean $\pm$ Std	78.95 $\pm$ 14.00	81.71 $\pm$ 14.43	87.57 $\pm$ 17.17	88.01 $\pm$ 17.44	61.91 $\pm$ 9.63	63.81 $\pm$ 8.63	66.96 $\pm$ 19.45	70.86 $\pm$ 17.98	54.35 $\pm$ 14.85	39.62 $\pm$ 12.54	68.50 $\pm$ 13.26	76.90 $\pm$ 10.19	54.60 $\pm$ 15.97	64.33 $\pm$ 11.97		

Table 4.7: Comparison of Detection and Generation Performance (P, Precision and R, Recall Values) on Full Length Videos for Various Deepfake Detection Models (ROWS) and Generators (COLUMNS). The bolded values highlight the best overall detection model across rows and the most challenging deepfake generator across columns.

Model	AniFaceDiff [11]		TPSMM [2]		FaceVid [1]		HyperReenact [14]		FaceFusion [12]		FADM [13]		SOTA [3, 4, 15, 16]		Detection Performance: Mean $\pm$ Std	
	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R
Capsule [41]	80.00	73.73	71.97	95.00	64.08	66.00	58.43	52.00	36.21	55.26	77.84	83.33	43.75	35.00	61.75 $\pm$ 16.78	65.76 $\pm$ 20.30
CORE [31]	67.00	77.01	93.27	97.00	81.58	31.00	89.55	60.00	56.25	23.68	91.25	46.79	87.50	35.00	80.91 $\pm$ 14.02	52.93 $\pm$ 26.67
EfficientNet-B4 [17]	69.00	76.67	88.17	82.00	76.09	35.00	50.00	11.00	42.11	21.05	83.58	35.90	100.00	35.00	72.71 $\pm$ 20.73	42.37 $\pm$ 26.88
FFD [42]	80.00	74.07	72.73	96.00	59.09	52.00	67.27	74.00	23.40	28.95	72.09	59.62	75.00	45.00	64.23 $\pm$ 19.15	61.38 $\pm$ 22.10
MesoNet [5]	100.00	68.03	51.55	100.00	51.55	100.00	51.55	100.00	28.79	100.00	62.40	100.00	50.00	100.00	56.55 $\pm$ 21.64	95.43 $\pm$ 12.08
MesoInception [5]	85.00	62.96	53.80	99.00	52.51	94.00	40.14	57.00	20.56	57.89	63.36	94.23	45.71	80.00	51.58 $\pm$ 19.95	77.87 $\pm$ 18.42
RECCE [18]	80.00	81.22	85.19	92.00	70.37	38.00	79.75	63.00	52.94	47.37	81.82	46.15	83.33	25.00	76.20 $\pm$ 11.29	56.11 $\pm$ 23.94
SRM [7]	86.00	75.44	68.18	90.00	53.85	49.00	66.67	84.00	40.00	73.68	71.62	67.95	53.85	35.00	62.88 $\pm$ 14.94	67.87 $\pm$ 19.48
UCF [43]	71.00	77.60	88.99	97.00	72.09	31.00	80.95	51.00	45.45	26.32	80.33	31.41	100.00	25.00	76.97 $\pm$ 17.10	48.48 $\pm$ 28.42
Xception [6]	76.00	76.38	81.15	99.00	69.33	52.00	77.67	80.00	47.73	55.26	81.60	65.38	84.62	55.00	74.01 $\pm$ 12.59	69.00 $\pm$ 17.14
Generation Performance: Mean $\pm$ Std	79.40 $\pm$ 9.69	74.31 $\pm$ 5.22	75.50 $\pm$ 14.54	94.70 $\pm$ 5.46	65.05 $\pm$ 10.50	54.80 $\pm$ 24.86	66.20 $\pm$ 15.97	63.20 $\pm$ 24.11	39.34 $\pm$ 12.09	48.95 $\pm$ 25.19	76.59 $\pm$ 9.15	63.08 $\pm$ 23.75	72.38 $\pm$ 22.13	47.00 $\pm$ 24.74		

Table 4.8: Comparison of Detection and Generation Performance (d' , Sensitivity Index and C, Decision Criterion Values) on Full Length Videos for Various Deepfake Detection Models (ROWS) and Generators (COLUMNS). The bolded values highlight the best overall detection model across rows and the most challenging deepfake generator across columns.

Model	AniFaceDiff [11]		TPSMM [2]		FaceVid [1]		HyperReenact [14]		FaceFusion [12]		FADM [13]		SOTA [3, 4, 15, 16]		Detection Performance: Mean $\pm$ Std	
	d'	C	d'	C	d'	C	d'	C	d'	C	d'	C	d'	C	d'	C
Capsule [41]	1.11	-0.29	1.91	-0.69	0.68	-0.07	0.32	0.11	0.40	0.07	1.24	-0.35	-0.26	0.26	0.77 $\pm$ 0.71	-0.14 $\pm$ 0.33
CORE [31]	1.88	0.50	3.32	-0.22	0.95	0.97	1.70	0.59	0.73	1.08	1.36	0.76	1.26	1.02	1.60 $\pm$ 0.86	0.67 $\pm$ 0.45
EfficientNet-B4 [17]	1.69	0.35	2.11	0.14	0.80	0.79	-0.04	1.21	0.39	1.00	0.83	0.78	5.98	3.37	1.68 $\pm$ 2.03	1.09 $\pm$ 1.07
FFD [42]	1.14	-0.27	2.05	-0.73	0.35	0.12	0.94	-0.17	-0.26	0.43	0.54	0.03	0.91	0.58	0.81 $\pm$ 0.72	0.00 $\pm$ 0.44
MesoNet [5]	0.00	-6.36	0.00	-6.36	0.00	-6.36	0.00	-6.36	0.00	-6.36	0.00	-6.36	0.00	-6.36	0.00 $\pm$ 0.00	-6.36 $\pm$ 0.00
MesoInception [5]	-0.27	-1.17	1.02	-1.82	0.25	-1.43	-1.13	-0.74	-1.11	-0.75	0.27	-1.44	-0.80	-1.24	-0.25 $\pm$ 0.81	-1.23 $\pm$ 0.39
RECCE [18]	1.75	0.04	2.36	-0.23	0.65	0.63	1.29	0.31	0.89	0.51	0.86	0.52	0.97	1.16	1.25 $\pm$ 0.61	0.42 $\pm$ 0.45
SRM [7]	1.21	-0.47	1.42	-0.57	0.11	0.08	1.13	-0.43	0.77	-0.25	0.6	-0.17	0.14	0.45	0.77 $\pm$ 0.52	-0.19 $\pm$ 0.36
UCF [43]	1.69	0.29	3.02	-0.37	0.64	0.82	1.16	0.56	0.50	0.89	0.65	0.81	5.69	3.52	1.91 $\pm$ 1.89	0.93 $\pm$ 1.22
Xception [6]	1.40	-0.01	3.02	-0.82	0.74	0.32	1.53	-0.08	0.82	0.28	1.09	0.15	1.41	0.58	1.43 $\pm$ 0.76	0.06 $\pm$ 0.45
Generation Performance: Mean $\pm$ Std	1.16 $\pm$ 0.74	-0.74 $\pm$ 2.03	2.02 $\pm$ 1.01	-1.17 $\pm$ 1.90	0.52 $\pm$ 0.32	-0.41 $\pm$ 2.20	0.69 $\pm$ 0.88	-0.50 $\pm$ 2.13	0.31 $\pm$ 0.62	-0.31 $\pm$ 2.20	0.74 $\pm$ 0.42	-0.53 $\pm$ 2.16	1.53 $\pm$ 2.38	0.33 $\pm$ 2.75		

We analyzed how participants’ prior experience with generative AI tools influenced their ability to detect deepfakes. Based on self-reported familiarity and usage of tools like ChatGPT, DALL·E, and MidJourney, participants were grouped into low, moderate, and high experience levels. As shown in Figure 4.3, detection error declined with increasing AI experience, with high-experience participants achieving the lowest error rates. A one-way ANOVA confirmed a significant effect of AI experience on discriminability ($p = 1.0137$), highlighting the value of hands-on exposure in improving media forensics skills. Further, as shown in Figure 4.4, low-experience participants tended to misclassify deepfakes as real, while high-experience individuals were more confident and accurate in identifying manipulated content. These findings suggest that familiarity with generative AI enhances both accuracy and confidence in synthetic media detection. This insight could inform the development of automated detection systems by highlighting the types of cues or heuristics used by experienced human observers. Incorporating human-like pattern recognition or training detection models using human error patterns may enhance robustness against increasingly sophisticated deepfakes.

Taken together, these findings offer critical insights into the strengths and limitations of both human and algorithmic deepfake detection systems. Humans consistently outperform current models across key metrics, including AUC, AP, d' , and precision, demonstrating superior accuracy, consistency, and robustness across variations in generative methods and video resolution. Notably, prior experience with AI plays a significant role: individuals with greater digital literacy not only make fewer errors but also exhibit stronger perceptual confidence when identifying manipulated content. While certain detection models such as FFD [42] and UCF [43] show potential, their performance remains inconsistent, revealing gaps in generalizability and interpretability. These insights highlight the need for developing detection systems that not only advance algorithmic performance but also incorporate human perceptual strengths, pointing toward a hybrid, human-in-the-loop approach as a promising direction in the ongoing battle against deepfake-driven misinformation.

4.3.3 Benchmarking and Performance Profiling of Detection Models

To extend the scope and depth of our research on deepfake detection, we performed a comprehensive comparative analysis of ten state-of-the-art detection models across seven distinct deepfake generation methods and the results of the experiments are shown in Tables 4.6, 4.7, 4.8. While our previous experiments were conducted on short 5-second video clips—common in existing literature—this study advances the field by evaluating model performance on full-length videos. This shift enables us to more accurately assess model behavior in realistic, temporally extended scenarios, where detection challenges such as motion consistency, long-term identity preservation, and temporal artifacts are more pronounced. Our evaluation spans a broad set of performance metrics, including AUC, average precision (AP), precision, recall, d' , and decision criterion (C), thereby offering a rich, multidimensional perspective on model capabilities. Furthermore,

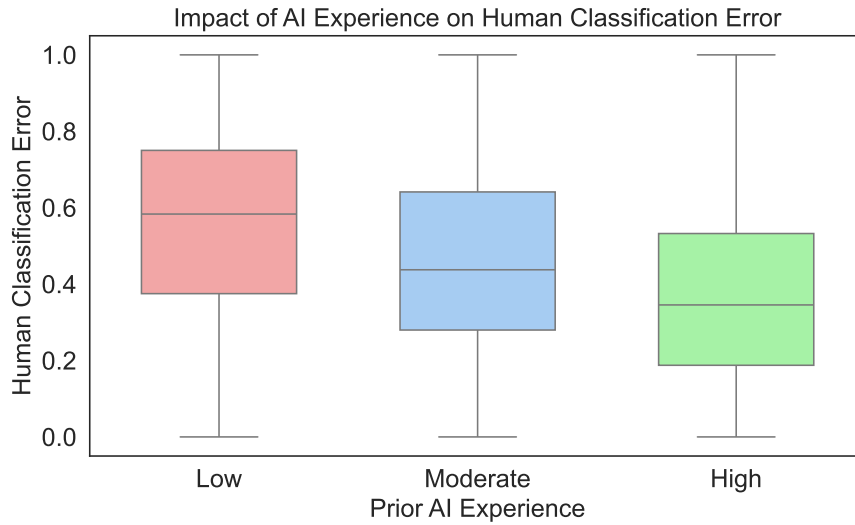


Figure 4.3: Aggregate impact of prior AI experience on fake video classification accuracy. This figure shows the average classification error across participant groups with different levels of AI experience. Participants with low AI experience had the highest overall error rates, while those with high experience performed significantly better. These results indicate a clear negative correlation between AI expertise and misclassification of fake videos at the group level.

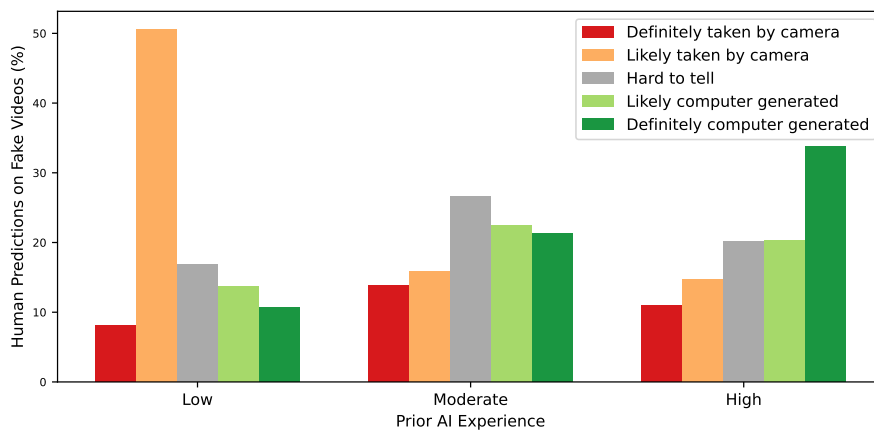


Figure 4.4: Individual classification responses to fake videos by AI experience level. This figure captures how individual participants rated individual fake videos, broken down by their level of prior AI experience. Participants with low AI experience were more likely to misclassify fake content as real, whereas those with high experience more frequently labeled videos as ‘Definitely computer generated.’ The results suggest that AI expertise improves detection accuracy on a video-by-video basis.

by incorporating complex and diverse generation methods such as AniFaceDiff [11], FaceFusion [12], and HyperReenact [14], we provide insights into generalization performance under varied manipulation paradigms. This study not only benchmarks model robustness in practical conditions but also informs the development of next-generation forensic tools capable of operating reliably in real-world media ecosystems.

Among the metrics reported, AUC and AP provide a threshold-independent view of performance. Xception [6], CORE [31], and EfficientNet-B4 [17] yield the highest AUCs across several generation methods, indicating reliable discrimination between real and fake samples. In contrast, models such as MesoNet [5] and FFD [42] exhibit relatively lower AUCs and AP values, especially on more challenging manipulations like FaceFusion [12] and HyperReenact [14], suggesting limited detection capability under complex generative distributions.

Precision and recall values further reveal model-specific trade-offs. Capsule [41] and RECCE [18] models show high precision but lower recall, reflecting a conservative detection strategy prone to fewer false positives but more false negatives. Conversely, models like UCF [43] and SRM [7] prioritize higher recall, potentially at the cost of increased false positives. These trends are substantiated by F1 scores, where Xception [6] and CORE [31] again lead, suggesting a favorable balance between precision and recall.

From a signal detection theory perspective, d' and C values offer insights into sensitivity and bias. Xception [6] and EfficientNet-B4 [17] consistently achieve high d' values across most generators, indicating strong separability between classes. MesoNet [5] and FFD [42] display both low d' and erratic C values, indicating susceptibility to both sensitivity loss and decision bias. Models with near-zero C values, like CORE [31] and EfficientNet-B4 [17], suggest an optimal trade-off between hit rate and false alarm rate.

Analyzing averages across generation methods reveals that certain models maintain stability, e.g., Xception [6], CORE [31], while others such as MesoNet [5] and FFD [42] experience significant performance degradation, especially under newer generation techniques like AniFaceDiff [11] and HyperReenact [14]. Standard deviation metrics further corroborate this, with high variance observed in MesoNet [5] and SRM [7], indicating model inconsistency across manipulation types. In contrast, the low standard deviation of Xception [6] across metrics and methods underscores its adaptability.

This extensive evaluation suggests that models pretrained on large-scale, diverse datasets with deep architectures, e.g., Xception [6], EfficientNet-B4 [17] generalize better to unseen or more complex manipulations. Simpler models, e.g., MesoNet [5], FFD [42] may overfit to specific patterns or fail to capture subtle semantic inconsistencies introduced by advanced generation pipelines. These findings emphasize the necessity for robustness and diversity in training data, as well as the use of a broad set of performance metrics to accurately assess detection systems in real-world forensic deployments.

4.3.4 Impact of Video Conditions on Detection Performance

Table 4.9 presents the average AUC (%) scores of deepfake detection models evaluated under varying input conditions, specifically comparing performance on full-length videos, 5-second clips, and 5-second clips at both low and high resolutions. These results reflect the mean detection performance of each model across multiple generative methods. Notably, human evaluators were only assessed on short clips and not on full-length videos due to practical limitations. Among automated models, Xception [6] and CORE [31] demonstrate the strongest overall performance, with Xception achieving the highest AUC on full-length inputs. This performance gain on full videos can be attributed to the richer temporal context and greater opportunity to observe forgery artifacts across frames, which enables models to extract more robust and representative features. Human detection accuracy improves noticeably with resolution, indicating that visual clarity significantly aids human judgment. In contrast, this resolution-dependent gain is not consistently reflected in automated models. For instance, MesoNet [5] and Meso4Inception [5] show substantial performance drops under low-resolution conditions, highlighting their sensitivity to input quality. On the other hand, RECCE [18] and UCF [43] display relatively stable AUC scores across all conditions, indicating robustness to both duration and resolution variations. These findings underscore the need for developing detection models that maintain high performance across a broad range of real-world input scenarios.

Table 4.9: Average AUC (%) of Deepfake Detection Models Across Varying Video Durations and Resolutions. Each value represents the model’s mean performance across multiple generative methods. Detection performance is evaluated on full-length videos, 5-second clips, and 5-second clips at low and high resolutions. The bolded values highlight the best detection model on each video condition.

Note: Human evaluation was not conducted on full-length videos; corresponding value is not available.

Detection Model	Full Length	5 Seconds	5s - Low Res	5s - High Res
Human	—	93.10	88.25	96.40
Capsule [41]	69.83	66.55	69.04	68.07
CORE [31]	77.56	68.74	71.50	71.93
EfficientNet-B4 [17]	63.34	64.24	68.04	64.79
FFD [42]	70.98	69.38	70.36	69.82
MesoNet [5]	47.70	49.73	47.61	51.89
Meso4Inception [5]	45.64	61.98	62.29	67.21
RECCE [18]	73.76	70.94	73.14	71.18
SRM [7]	69.94	63.43	64.11	63.89
UCF [43]	73.49	72.49	72.29	74.00
Xception [6]	83.24	68.43	64.57	72.89

4.3.5 Multimodal Fusion Strategies and Detection Results on Novel Scenarios

To systematically evaluate the effectiveness of unimodal and multimodal approaches for deepfake detection, we conducted extensive experiments using a balanced training set comprising 6,500 real and 6,500 fake samples from the FakeAVCeleb dataset. For evaluation, a held-out test set containing 2,500 real and 2,500 fake samples was constructed using the AV-Deepfake1M dataset [66], which includes audio-visual deepfake videos of 2,068 subjects featuring realistic and diverse cross-modal manipulations generated through ChatGPT-based transcript editing and high-fidelity synthesis.

To simulate challenging real-world conditions, we evaluated each model under two conditions: full-length videos and 5-second video clips. Furthermore, we explored the impact of degraded resolution on detection performance. Low-quality videos were generated by downsampling high-resolution frames to 128×128 using bicubic interpolation, then re-encoding them at lower bitrates (CRF=35) using FFmpeg. Low-quality audio was created by downsampling the original waveform to 8kHz mono-channel and applying low-pass filtering to reduce spectral richness. These manipulations ensured that performance under constrained visual and auditory conditions could be reliably assessed.

Table 4.10 presents results across multiple configurations. We report Area Under the ROC Curve (AUC), Average Precision (AP), Precision, Recall, the discriminability index (d'), and decision bias (C). These metrics collectively provide a holistic view of model efficacy.

Among unimodal baselines, the audio-only model consistently outperformed the video-only model across all metrics. For full-length inputs, the audio-based model achieved an AUC of 78.5% and $d' = 1.92$, compared to the video-based model’s 64.5% AUC and $d' = 0.65$. This trend persisted in 5-second evaluations, further underscoring the discriminative advantage of speech signals in deepfake detection, particularly under resolution-constrained settings.

Multimodal fusion strategies exhibited consistent performance improvements over unimodal counterparts. For full-length inputs, Early Fusion achieved the highest overall performance (AUC = 83.0%, $d' = 2.11$), suggesting that learning joint audio-visual representations early in the network effectively captures complementary cues. Cross-Modal Attention also performed competitively (AUC = 81.0%, $d' = 2.02$), particularly excelling in low-resolution conditions. Its robustness in 5-second clips further demonstrates the strength of attention mechanisms in aligning modalities under limited context.

Interestingly, under low video resolution and low audio quality, Cross-Modal Attention outperformed all other fusion strategies in the 5-second condition (AUC = 67.0%, $d' = 1.45$), suggesting superior resilience when perceptual quality is degraded. In contrast, Late Fusion consistently underperformed across both durations and resolutions (e.g., 5-second AUC = 63.1% under low-low condition), highlighting its limitations in modeling cross-modal correlations when audio and video are processed independently.

Table 4.10: Performance comparison of unimodal and multimodal deepfake detection across varying durations, resolutions, and fusion strategies. The table reports AUC, AP, precision, recall, Sensitivity Index (d-prime), and Decision Criterion (C) for both full-length and 5-second video clips. Unimodal (video-only or audio-only) and multimodal (audio-video) configurations are evaluated under high/low resolution conditions. Multimodal fusion strategies include Early Fusion, Late Fusion, and Cross-Modal Attention. While all models show performance degradation on shorter clips, Cross-Modal Attention consistently outperforms other fusion approaches under low-resolution settings, highlighting its effectiveness in learning cross-modal representations when visual quality is limited.

Duration	Modality	Video Resolution	Audio Resolution	Fusion Strategy	AUC (%)	AP (%)	Precision (%)	Recall (%)	d'	C
Full Length	Video	High	N/A	N/A	64.5	66.32	65.53	53	0.65	0.25
		Low			58	60.1	60.75	48	0.52	0.3
	Audio	N/A	High	N/A	78.5	77.5	86.14	78.36	1.92	0.18
			Low		71.2	70	80.75	72.1	1.55	0.22
	Video & Audio	High	High	Early Fusion	83	82.5	88.62	80.61	2.11	0.19
				Late Fusion	75.5	74	77.58	82.87	1.64	-0.13
				Cross-Modal Attention	81	80.5	87.47	79.48	2.02	0.19
	Video & Audio	High	Low	Early Fusion	79.2	77.1	84.2	77.6	1.88	0.17
				Late Fusion	73	71.2	75.5	78.9	1.51	-0.09
				Cross-Modal Attention	77.5	76.3	83.25	76.4	1.85	0.15
	Video & Audio	Low	High	Early Fusion	76.8	75	83.7	75.1	1.79	0.14
				Late Fusion	70.3	69.2	72.44	77.5	1.43	-0.12
				Cross-Modal Attention	74.2	73.6	81.35	73.9	1.74	0.13
	Video & Audio	Low	Low	Early Fusion	70	68.3	76.15	70.6	1.42	0.08
				Late Fusion	67.8	65.9	70.9	73.5	1.31	-0.1
				Cross-Modal Attention	72	71	78.8	72.8	1.61	0.12
5-second	Video	High	N/A	N/A	60	61.7	61.6	48.76	0.59	0.23
		Low			53.9	55.9	57.1	44.16	0.47	0.27
	Audio	N/A	High	N/A	73	72.1	80.97	72.09	1.73	0.16
			Low		66.2	65.1	75.91	66.33	1.4	0.2
	Video & Audio	High	High	Early Fusion	77.2	76.7	83.3	74.16	1.9	0.17
				Late Fusion	70.2	68.8	72.93	76.24	1.48	-0.12
				Cross-Modal Attention	75.3	74.9	82.22	73.12	1.82	0.17
	Video & Audio	High	Low	Early Fusion	73.7	71.7	79.15	71.39	1.69	0.15
				Late Fusion	67.9	66.2	70.97	72.59	1.36	-0.08
				Cross-Modal Attention	72.1	71	78.25	70.29	1.67	0.14
	Video & Audio	Low	High	Early Fusion	71.4	69.8	78.68	69.09	1.61	0.13
				Late Fusion	65.4	64.4	68.09	71.3	1.29	-0.11
				Cross-Modal Attention	69	68.4	76.47	67.99	1.57	0.12
	Video & Audio	Low	Low	Early Fusion	65.1	63.5	71.58	64.95	1.28	0.07
				Late Fusion	63.1	61.3	66.65	67.62	1.18	-0.09
				Cross-Modal Attention	67	66	74.07	66.98	1.45	0.11

In summary, our findings confirm that integrating auditory and visual modalities substantially enhances detection robustness—especially when signal quality is compromised. Among fusion techniques, Early Fusion remains the most effective overall, while Cross-Modal Attention shows superior adaptability in low-resource or short-duration scenarios.

4.4 Summary

This chapter presented a comprehensive evaluation of deepfake detection models across multiple experimental settings, including variations in video resolution, clip duration, and input modalities. Benchmarking results showed that unimodal models varied in performance depending on the manipulation technique and input quality, with high-resolution and longer videos generally yielding better results. Human evaluators outperformed automated detectors, especially against state-of-the-art deepfakes, with prior AI experience further enhancing detection accuracy. Due to computational constraints, full-length videos were processed in 2.56-second chunks with zero padding and averaged predictions. The proposed multimodal framework, combining video and audio inputs using pretrained backbones and fusion strategies, consistently outperformed unimodal baselines. Among the

fusion techniques, early fusion achieved the best overall performance, while cross-modal attention proved especially effective in low-resolution scenarios. These findings underscore the importance of multimodal integration in building robust and generalizable deepfake detection systems.

Chapter 5

Conclusion and Future Directions

This chapter concludes the study by summarizing the key findings, discussing the limitations encountered during the research, and outlining potential future directions. We reflect on the results obtained from the implementation, benchmarking, human evaluation, and multimodal integration of deepfake detection models.

5.1 Summary

Our study establishes a rigorous benchmark that reveals a growing gap between the sophistication of modern deepfake generation and the capabilities of current detection methods. Despite notable advances in detection, even state-of-the-art models struggle against deepfakes produced by cutting-edge synthesis techniques—particularly those based on diffusion models and Neural Radiance Fields (NeRF). Human evaluators, while effective in detecting certain manipulations, also face significant challenges with the most realistic and high-quality forgeries. These findings underscore the urgent need for more resilient and adaptive detection strategies. By benchmarking both human and machine performance, our study offers critical insights into the limitations of existing approaches and lays the groundwork for developing next-generation detection systems that integrate human perceptual strengths with automated techniques.

5.2 Limitation

Despite the comprehensive nature of this research, several limitations were identified. One of the major challenges was the high computational resources required for training and evaluating deepfake detection models, especially when dealing with longer videos, which demand more GPU support and RAM. Additionally, there is a lack of publicly available datasets specifically focusing on multimodal talking head deepfakes, limiting the scope of evaluation. Another limitation arises from the fact that not all deepfake generation models are open source due to ethical concerns, making it difficult to generate synthetic content for research purposes. Addressing these limitations in future research would help enhance

the generalizability and robustness of deepfake detection techniques.

5.3 Future Work

In future research, we plan to expand the scope of our multimodal deepfake detection framework in two significant ways. First, we aim to incorporate the remaining two modality configurations—Fake Video with Real Audio and Real Video with Fake Audio—to study how partial manipulation affects detection performance. Including these nuanced combinations will allow for a more comprehensive evaluation of the model’s ability to capture asymmetric inconsistencies between modalities.

Second, we intend to explore advanced fusion techniques and integrate additional modalities (e.g., text or physiological cues) to further improve the robustness and generalizability of deepfake detection systems. Another important direction is the inclusion of a human evaluation study, in which participants will be asked to detect multimodal deepfakes. This will help us understand how humans process audio-visual inconsistencies and identify potential perceptual cues that can inform future model design. By bridging the gap between human and machine detection capabilities, this research will contribute to building more effective and interpretable detection frameworks.

References

- [1] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10039–10049, 2021.
- [2] Jian Zhao and Hui Zhang. Thin-plate spline motion model for image animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3657–3666, 2022.
- [3] Sicheng Xu, Guojun Chen, Yu-Xiao Guo, Jiaolong Yang, Chong Li, Zhenyu Zang, Yizhong Zhang, Xin Tong, and Baining Guo. Vasa-1: Lifelike audio-driven talking faces generated in real time. *Advances in Neural Information Processing Systems*, 37:660–684, 2024.
- [4] Ziqiao Peng, Wentao Hu, Yue Shi, Xiangyu Zhu, Xiaomei Zhang, Hao Zhao, Jun He, Hongyan Liu, and Zhaoxin Fan. Synctalk: The devil is in the synchronization for talking head synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 666–676, 2024.
- [5] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Mesonet: a compact facial video forgery detection network. In *2018 IEEE international workshop on information forensics and security (WIFS)*, pages 1–7. IEEE, 2018.
- [6] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1–11, 2019.
- [7] Yuchen Luo, Yong Zhang, Junchi Yan, and Wei Liu. Generalizing face forgery detection with high-frequency features. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16317–16326, 2021.
- [8] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28130–28139, 2024.

- [9] Alexander Groshev, Anastasia Maltseva, Daniil Chesakov, Andrey Kuznetsov, and Denis Dimitrov. Ghost—a new face swap approach for image and video domains. *IEEE Access*, 10:83452–83462, 2022.
- [10] Stella Bounareli, Christos Tzelepis, Vasileios Argyriou, Ioannis Patras, and Georgios Tzimiropoulos. Stylemask: Disentangling the style space of stylegan2 for neural face reenactment. In *2023 IEEE 17th international conference on automatic face and gesture recognition (FG)*, pages 1–8. IEEE, 2023.
- [11] Ken Chen, Sachith Seneviratne, Wei Wang, Dongting Hu, Sanjay Saha, Md Tarek Hasan, Sanka Rasnayaka, Tamasha Malepathirana, Mingming Gong, and Saman Halgamuge. Anifacediff: High-fidelity face reenactment via facial parametric conditioned diffusion models. *arXiv preprint arXiv:2406.13272*, 2024.
- [12] FaceFusion. Facefusion: A github repository for next generation face swapper and enhancer, 2024. Accessed: 2024-08-14.
- [13] Bohan Zeng, Xuhui Liu, Sicheng Gao, Boyu Liu, Hong Li, Jianzhuang Liu, and Baochang Zhang. Face animation with an attribute-guided diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 628–637, 2023.
- [14] Stella Bounareli, Christos Tzelepis, Vasileios Argyriou, Ioannis Patras, and Georgios Tzimiropoulos. Hyperreenact: one-shot reenactment via jointly learning to refine and retarget faces. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7149–7159, 2023.
- [15] Yifeng Ma, Shiwei Zhang, Jiayu Wang, Xiang Wang, Yingya Zhang, and Zhidong Deng. Dreamtalk: When expressive talking head generation meets diffusion probabilistic models. *arXiv preprint arXiv:2312.09767*, 2023.
- [16] Michał Stypułkowski, Konstantinos Vougioukas, Sen He, Maciej Zikeba, Stavros Petridis, and Maja Pantic. Diffused heads: Diffusion models beat gans on talking-face generation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5091–5100, 2024.
- [17] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.
- [18] Junyi Cao, Chao Ma, Taiping Yao, Shen Chen, Shouhong Ding, and Xiaokang Yang. End-to-end reconstruction-classification learning for face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4113–4122, 2022.

- [19] Rami Mubarak, Tariq Alsboui, Omar Alshaikh, Isa Inuwa-Dutse, Saad Khan, and Simon Parkinson. A survey on the detection and impacts of deepfakes in visual, audio, and textual formats. *Ieee Access*, 11:144497–144529, 2023.
- [20] Saima Waseem, Syed Abdul Rahman Syed Abu Bakar, Bilal Ashfaq Ahmed, Zaid Omar, and Taiseer Abdalla Elfadil Eisa. Deepfake on face and expression swap: A review. *IEEE Access*, 11:117865–117906, 2023.
- [21] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.
- [22] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015.
- [23] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.
- [24] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.
- [25] Ashifur Rahman, Md Mazharul Islam, Mohasina Jannat Moon, Tahera Tasnim, Nipo Siddique, Md Shahiduzzaman, and Samsuddin Ahmed. A qualitative survey on deep learning based deep fake video creation and detection method. *Aust. J. Eng. Innov. Technol*, 4(1):13–26, 2022.
- [26] Gan Pei, Jiangning Zhang, Menghan Hu, Zhenyu Zhang, Chengjie Wang, Yunsheng Wu, Guangtao Zhai, Jian Yang, Chunhua Shen, and Dacheng Tao. Deepfake generation and detection: A benchmark and survey. *arXiv preprint arXiv:2403.17881*, 2024.
- [27] Arash Heidari, Nima Jafari Navimipour, Hasan Dag, and Mehmet Unal. Deepfake detection using deep learning methods: A systematic and comprehensive review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 14(2):e1520, 2024.
- [28] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.
- [29] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

- [30] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017.
- [31] Yunsheng Ni, Depu Meng, Changqian Yu, Chengbin Quan, Dongchun Ren, and Youjian Zhao. Core: Consistent representation learning for face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12–21, 2022.
- [32] Leandro A Passos, Danilo Jodas, Kelton AP Costa, Luis A Souza Júnior, Douglas Rodrigues, Javier Del Ser, David Camacho, and João Paulo Papa. A review of deep learning-based approaches for deepfake content detection. *Expert Systems*, 41(8):e13570, 2024.
- [33] Tianyi Wang, Xin Liao, Kam Pui Chow, Xiaodong Lin, and Yinglong Wang. Deepfake detection: A comprehensive survey from the reliability perspective. *ACM Computing Surveys*, 57(3):1–35, 2024.
- [34] Felix Juefei-Xu, Run Wang, Yihao Huang, Qing Guo, Lei Ma, and Yang Liu. Countering malicious deepfakes: Survey, battleground, and horizon. *International journal of computer vision*, 130(7):1678–1734, 2022.
- [35] Lucas Whittaker, Rory Mulcahy, Kate Letheren, Jan Kietzmann, and Rebekah Russell-Bennett. Mapping the deepfake landscape for innovation: A multidisciplinary systematic review and future research agenda. *Technovation*, 125:102784, 2023.
- [36] Sergi D Bray, Shane D Johnson, and Bennett Kleinberg. Testing human ability to detect ‘deepfake’ images of human faces. *Journal of Cybersecurity*, 9(1):tyad011, 2023.
- [37] Davide Salvi, Honggu Liu, Sara Mandelli, Paolo Bestagini, Wenbo Zhou, Weiming Zhang, and Stefano Tubaro. A robust approach to multimodal deepfake detection. *Journal of Imaging*, 9(6):122, 2023.
- [38] Michael Lomnitz, Zigfried Hampel-Arias, Vishal Sandesara, and Simon Hu. Multimodal approach for deepfake detection. In *2020 IEEE applied imagery pattern recognition workshop (AIPR)*, pages 1–9. IEEE, 2020.
- [39] Xiaolong Liu, Yang Yu, Xiaolong Li, and Yao Zhao. Mcl: multimodal contrastive learning for deepfake detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(4):2803–2813, 2023.
- [40] Muhammad Anas Raza and Khalid Mahmood Malik. Multimodaltrace: Deepfake detection using audiovisual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 993–1000, 2023.

- [41] Huy H Nguyen, Junichi Yamagishi, and Isao Echizen. Capsule-forensics: Using capsule networks to detect forged images and videos. In *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 2307–2311. IEEE, 2019.
- [42] Hao Dang, Feng Liu, Joel Stehouwer, Xiaoming Liu, and Anil K Jain. On the detection of digital face manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern recognition*, pages 5781–5790, 2020.
- [43] Zhiyuan Yan, Yong Zhang, Yanbo Fan, and Baoyuan Wu. Ucf: Uncovering common features for generalizable deepfake detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22412–22423, 2023.
- [44] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2387–2395, 2016.
- [45] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.
- [46] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics: A large-scale video dataset for forgery detection in human faces. *arXiv preprint arXiv:1803.09179*, 2018.
- [47] Karen Simonyan. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [48] Qualtrics. Qualtrics survey platform. <https://www.qualtrics.com/lp/survey-platform/>, 2025. Accessed: 2025-05-05.
- [49] Prolific. Prolific. <https://www.prolific.com/>. Accessed: 2025-05-06.
- [50] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, volume 2, page 4, 2021.
- [51] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.
- [52] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4489–4497, 2015.

- [53] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.
- [54] Xiankai Huang and Zhibin Cai. A review of video action recognition based on 3d convolution. *Computers and Electrical Engineering*, 108:108713, 2023.
- [55] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.
- [56] Hasam Khalid, Shahroz Tariq, Minha Kim, and Simon S Woo. Fakeavceleb: A novel audio-video multimodal deepfake dataset. *arXiv preprint arXiv:2108.05080*, 2021.
- [57] Kunlin Liu, Ivan Perov, Daiheng Gao, Nikolay Chervoniy, Wenbo Zhou, and Weiming Zhang. Deepfacelab: Integrated, flexible and extensible face-swapping framework. *Pattern Recognition*, 141:109628, 2023.
- [58] Yuval Nirkin, Yosi Keller, and Tal Hassner. Fsgan: Subject agnostic face swapping and reenactment. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7184–7193, 2019.
- [59] Xuanhong Chen, Bingbing Ni, Yutian Liu, Naiyuan Liu, Zhilin Zeng, and Hang Wang. Simswap++: Towards faster and high-quality identity swapping. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(1):576–592, 2023.
- [60] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. Faceshifter: Towards high fidelity and occlusion aware face swapping. *arXiv preprint arXiv:1912.13457*, 2019.
- [61] Donald G Childers, Ke Wu, DM Hicks, and B Yegnanarayana. Voice conversion. *Speech Communication*, 8(2):147–158, 1989.
- [62] Paul Taylor. *Text-to-speech synthesis*. Cambridge university press, 2009.
- [63] Kaizhi Qian, Yang Zhang, Shiyu Chang, Xuesong Yang, and Mark Hasegawa-Johnson. Autovc: Zero-shot voice style transfer with only autoencoder loss. In *International Conference on Machine Learning*, pages 5210–5219. PMLR, 2019.
- [64] Isaac Elias, Heiga Zen, Jonathan Shen, Yu Zhang, Ye Jia, RJ Skerry-Ryan, and Yonghui Wu. Parallel tacotron 2: A non-autoregressive neural tts model with differentiable duration modeling. *arXiv preprint arXiv:2103.14574*, 2021.
- [65] Kundan Kumar, Rithesh Kumar, Thibault De Boissiere, Lucas Gestein, Wei Zhen Teoh, Jose Sotelo, Alexandre De Brebisson, Yoshua Bengio, and Aaron C Courville. Melgan: Generative adversarial networks for conditional waveform synthesis. *Advances in neural information processing systems*, 32, 2019.

-
- [66] Zhixi Cai, Abhinav Dhall, Shreya Ghosh, Munawar Hayat, Dimitrios Kollias, Kalin Stefanov, and Usman Tariq. 1m-deepfakes detection challenge. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 11355–11359, 2024.
 - [67] Charles X Ling, Jin Huang, Harry Zhang, et al. Auc: a statistically consistent and more discriminating measure than accuracy. In *Ijcai*, volume 3, pages 519–524, 2003.
 - [68] David Marvin Green, John A Swets, et al. *Signal detection theory and psychophysics*, volume 1. Wiley New York, 1966.